Virtualization and Cloud Computing in Support of Science at Fermilab

Keith Chadwick

Grid & Cloud Computing Department

Fermilab

Work supported by the U.S. Department of Energy under contract No. DE-AC02-07CH11359

Fermilab

My Background

Currently:

- 26 year employee of Fermilab,
- Grid and Cloud Computing Department Head.

Previously:

- Fermilab Campus Grid Services Project Leader,
- Managed various large ($500K to $1M) acquisitions,
- Data Communications Department Deputy Head,
- Fiber to the Desktop Project Manager,
- Distributed Software Group Leader,
- VMS Systems Administrator/Developer,
- CDF Experiment offline code manager,
- "Postdoc" on the CDF Experiment.

Outline

About Fermilab and Science at Fermilab

- Three frontiers of science

The beginnings

- When rocks were soft…

Fermilab Campus Grid (FermiGrid)

- Highly available virtualized services

FermiCloud

- Highly available IaaS cloud services

General Physics Compute Facility (GPCF)

- Statically deployed virtualized compute services

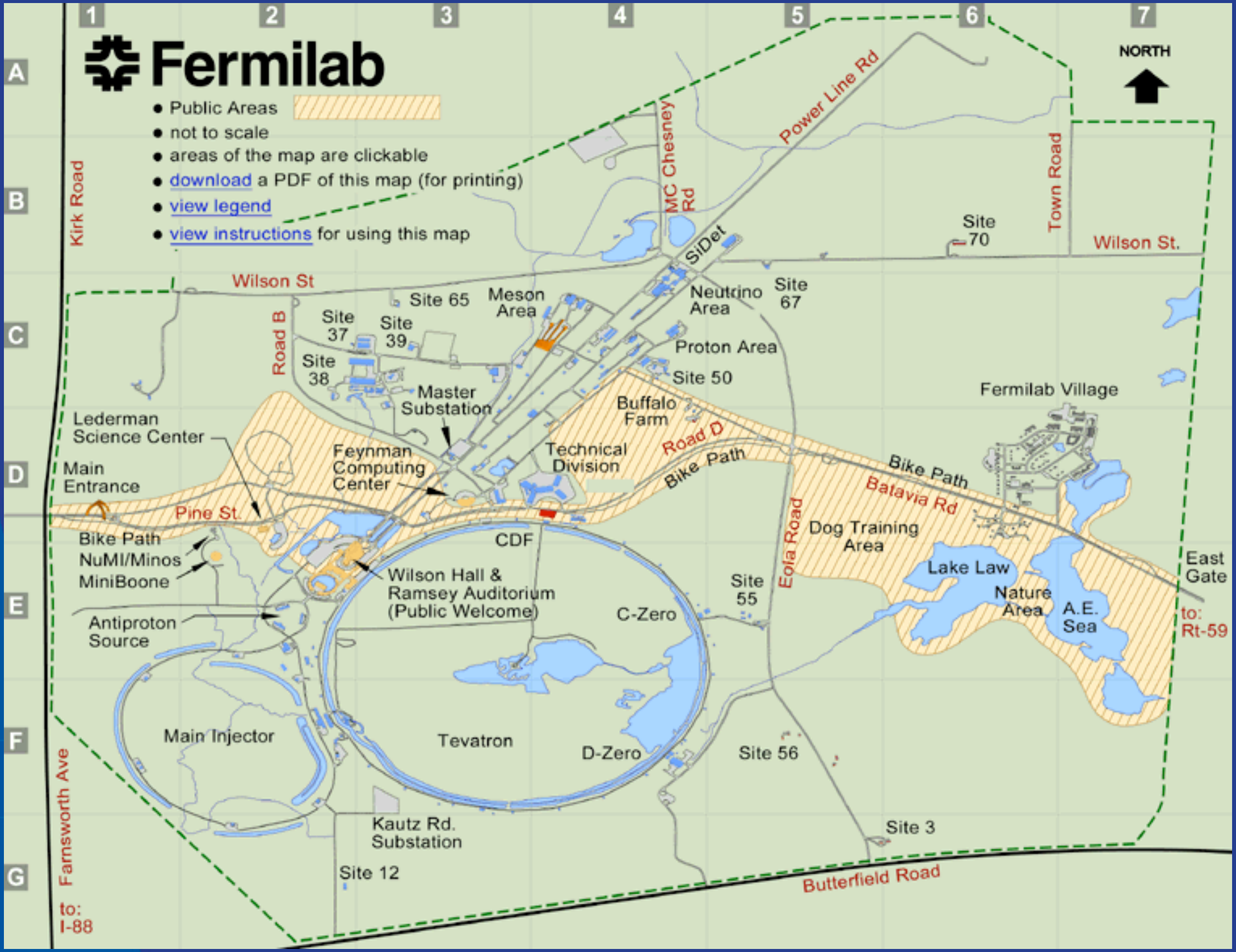
VMware virtual services

- Business critical virtualized services

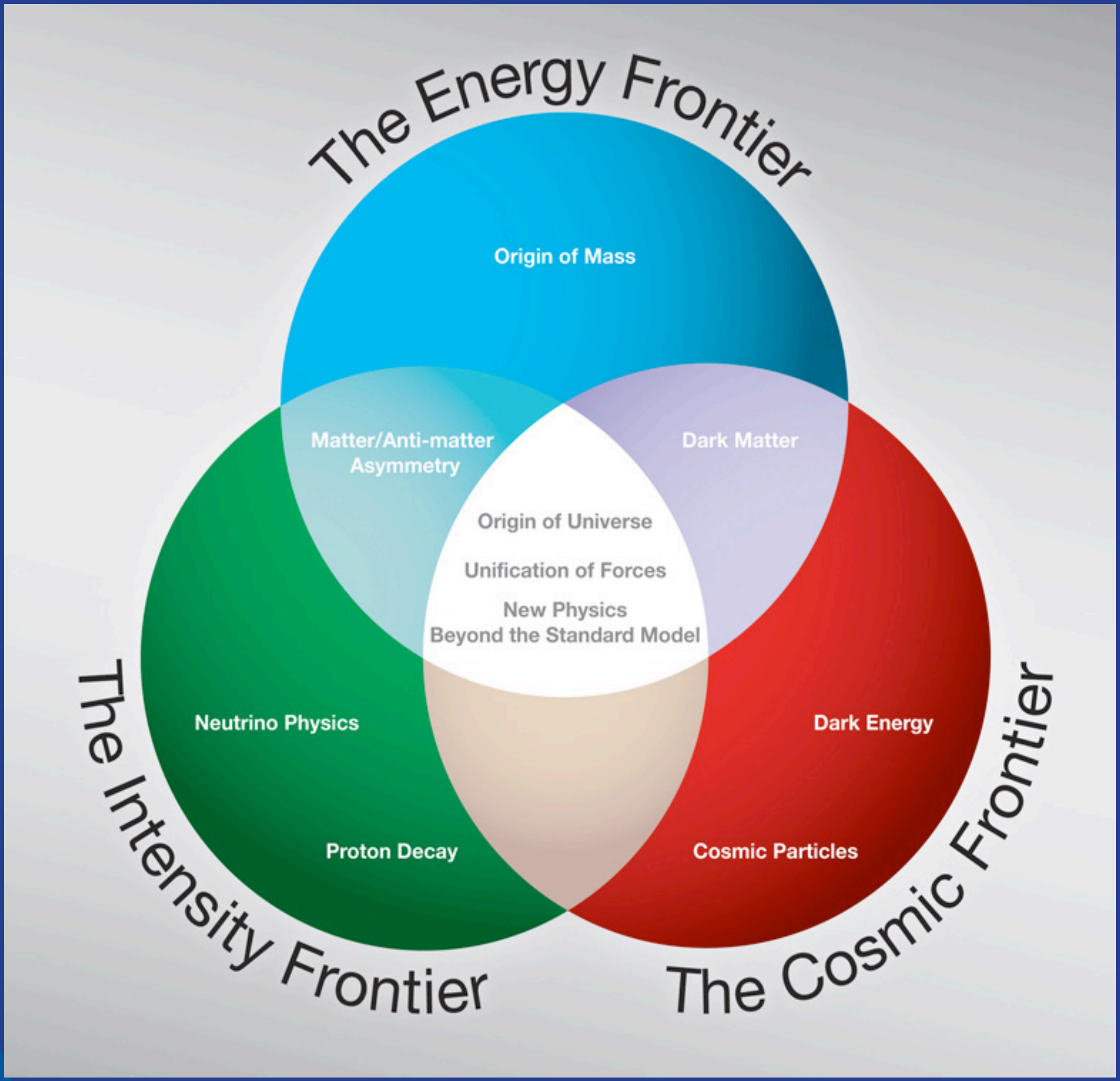
Plans for the future

About Fermilab

Fermilab Site
(~50 miles west of downtown Chicago)



Three Frontiers of Science



Energy, Intensity, Cosmic

Energy Frontier:

- At the Energy Frontier, scientists build advanced particle accelerators to explore the fundamental constituents and architecture of the universe. There they expect to encounter new phenomena not seen since the immediate aftermath of the big bang. Subatomic collisions at the energy frontier will produce particles that signal these new phenomena, from the origin of mass to the existence of extra dimensions.

Intensity Frontier:

- At the Intensity Frontier, scientists use accelerators to create intense beams of trillions of particles for neutrino experiments and measurements of ultra-rare processes in nature. Measurements of the mass and other properties of the neutrinos are key to the understanding of new physics beyond today's models and have critical implications for the evolution of the universe. Precise observations of rare processes provide a way to explore high energies, providing an alternate, powerful window to the nature of fundamental interactions.

Cosmic Frontier:

- At the Cosmic Frontier, astrophysicists use the cosmos as a laboratory to investigate the fundamental laws of physics from a perspective that complements experiments at particle accelerators. Thus far, astrophysical observations, including the bending of light known as gravitational lensing and the properties of supernovae, reveal a universe consisting mostly of dark matter and dark energy. A combination of underground experiments and telescopes, both ground- and space-based, will explore these mysterious dark phenomena that constitute 95 percent of the universe.

Fermilab Computing Facilities

Feynman Computing Center (FCC):

FCC-2 Computer Room,

FCC-3 Computer Room(s),

All rooms have refrigerant based cooling, UPS and Generators.

Grid Computing Center (GCC):

GCC-A, GCC-B, GCC-C Computer Rooms,

GCC-TRR Tape Robot Room,

GCC-Network-A, GCC-Network-B,

All rooms have refrigerant based cooling, UPS and "quick connect" taps for Generators.

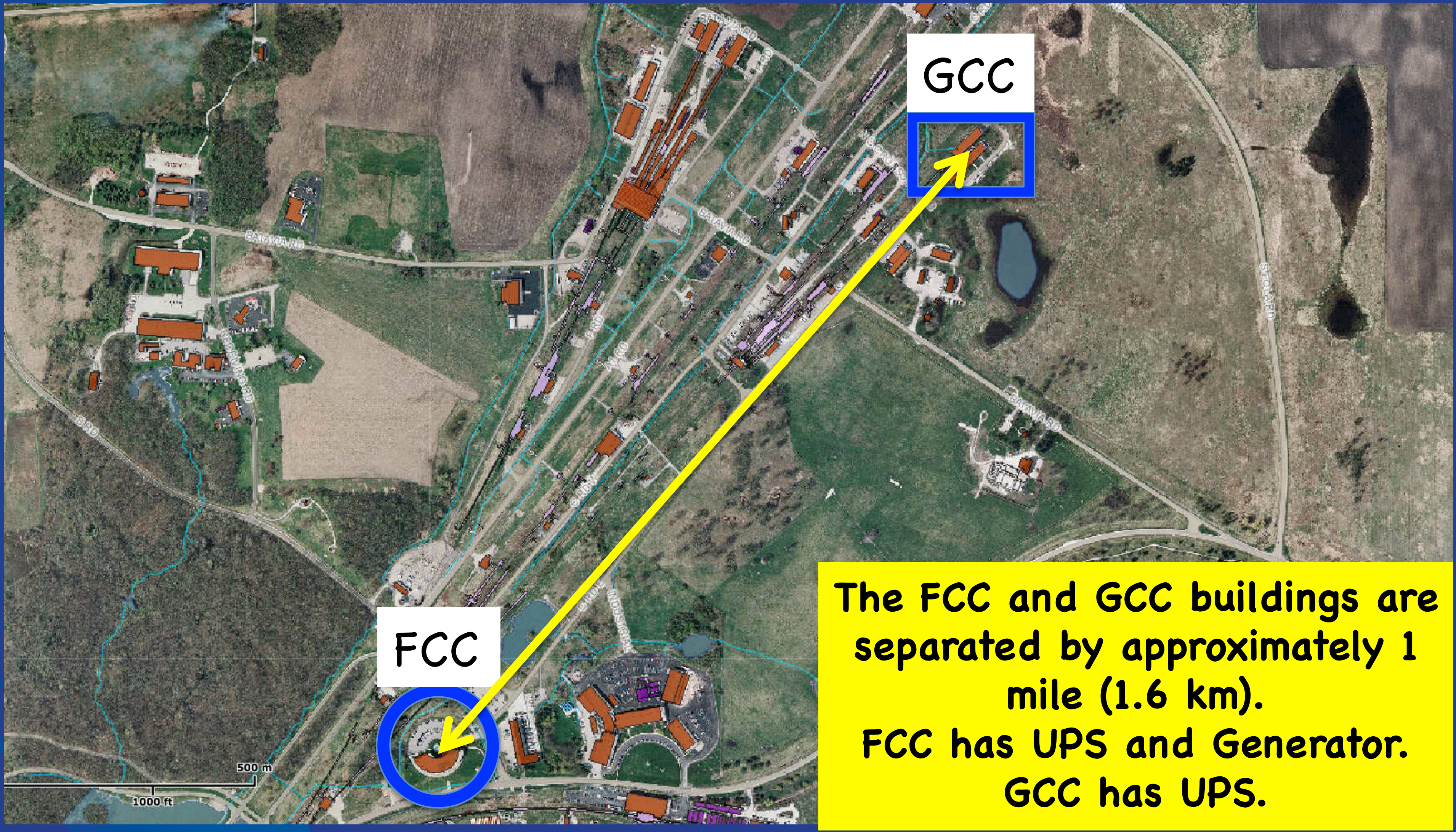
Lattice Computing Center (LCC):

LCC-107 & LCC-108 Computer Room,

Refrigerant based cooling,

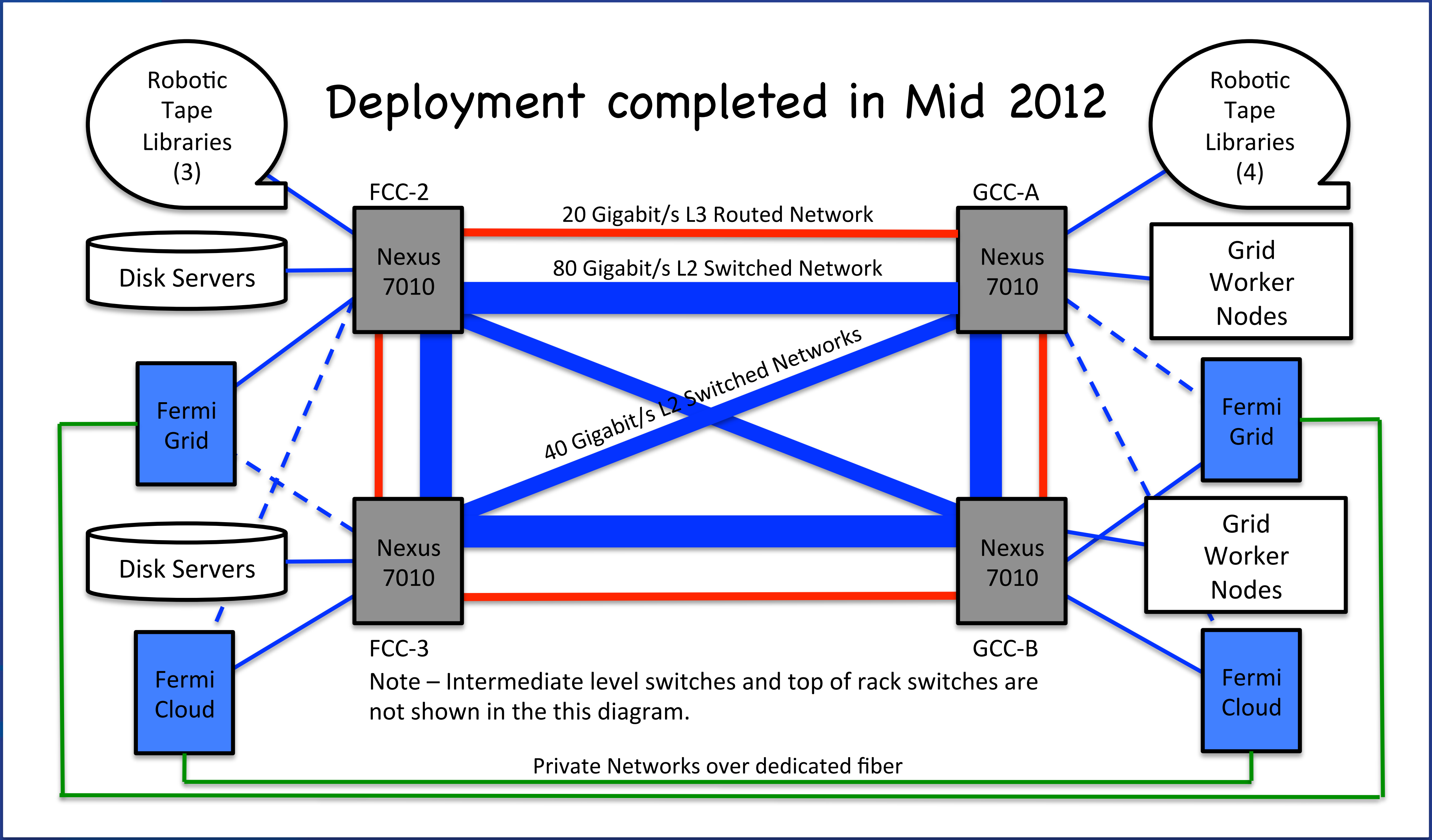
No UPS or Generator.

FCC and GCC



The FCC and GCC buildings are separated by approximately 1 mile (1.6 km).
FCC has UPS and Generator.
GCC has UPS.

Distributed Network Core Provides Redundant Connectivity


Deployment completed in Mid 2012
Robotic Tape Libraries (3)
Robotic Tape Libraries (4)
FCC-2
GCC-A
20 Gigabit/s L3 Routed Network
80 Gigabit/s L2 Switched Network
Nexus 7010
Nexus 7010
Disk Servers
Grid Worker Nodes
40 Gigabit/s L2 Switched Networks
Fermi Grid
Fermi Grid
Nexus 7010
Nexus 7010
Disk Servers
Grid Worker Nodes
Fermi Cloud
Fermi Cloud
FCC-3
GCC-B
Private Networks over dedicated fiber

Note – Intermediate level switches and top of rack switches are not shown in the this diagram.

Fermilab Campus Grid (FermiGrid), FermiGrid-HA, FermiGrid-HA2

FermiGrid – A “Quick” History

In 2004, the Fermilab Computing Division undertook the strategy of placing all of its production resources in a Grid "meta-facility" infrastructure called FermiGrid (the Fermilab Campus Grid).

In April 2005, the first “core” FermiGrid services were deployed.

In 2007, the FermiGrid-HA project was commissioned with the goal of delivering 99.999% core service availability (not including building or network outages).

During 2008, all of the FermiGrid services were redeployed in Xen virtual machines.

What Environment Do We Operate In?

Risk	Mitigation
Service Failure	Redundant copies of Service (with service monitoring).
System Failure	Redundant copies of Service (on separate systems with service monitoring).
Network Failure	Redundant copies of Service (on separate systems with independent network connections and service monitoring).
Building Failure	Redundant copies of Service (on separate systems with independent network connections in separate buildings and service monitoring).

FermiGrid-HA – Deployed in 2007

FermiGrid-HA2 – Deployed in 2011

FermiGrid-HA/FermiGrid-HA2

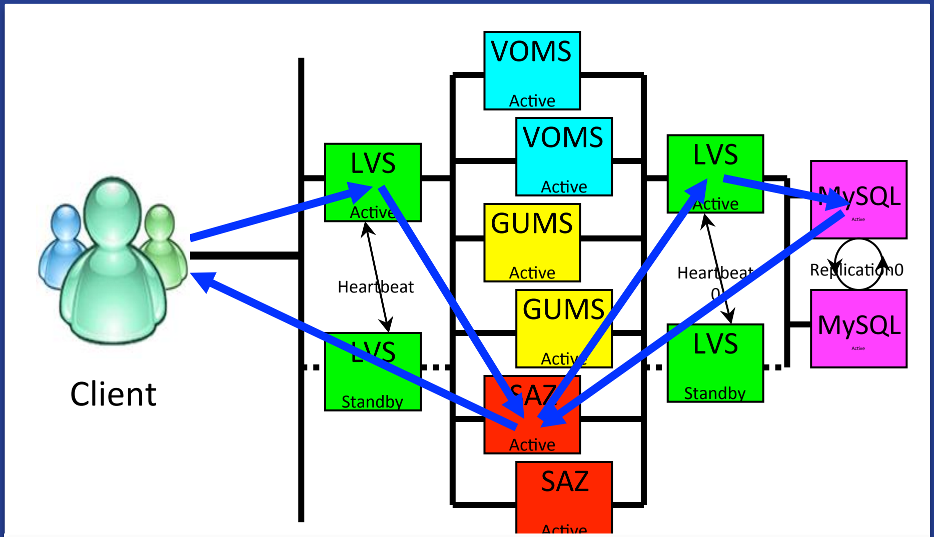
FermiGrid-HA uses three key technologies:

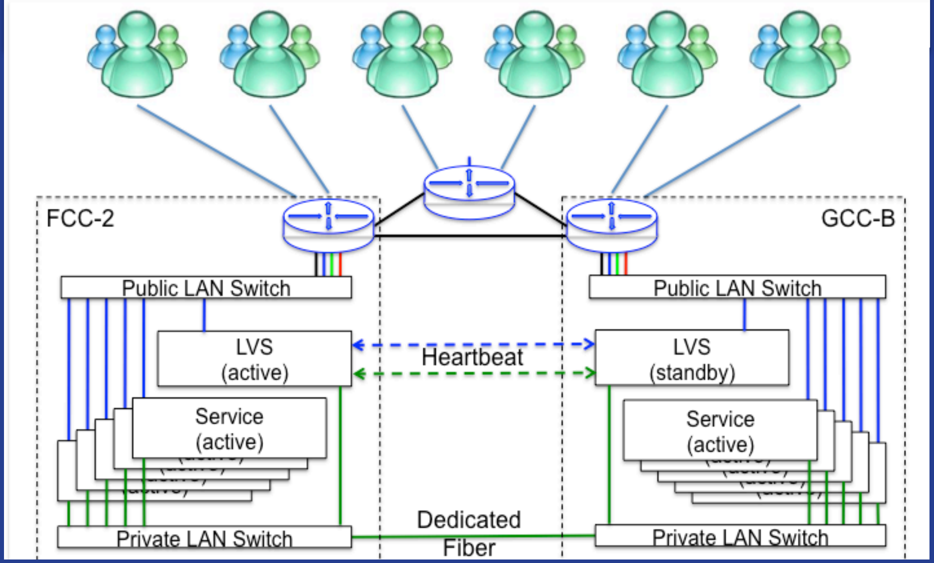
- Linux Virtual Server (LVS),
- Xen Hypervisor,
- MySQL Circular Replication.

FermiGrid-HA2 added:

- Redundant services in both FCC-2 and GCC-B,
- Non-redundant services are split across both locations, and go to reduced capacity in the event of building or network outage.

Deployment has been tested under real world conditions.





FermiGrid-HA2 Experience

In 2009, based on operational experience and plans for redevelopment of the FCC-1 computer room, the FermiGrid-HA2 project was established to split the set of FermiGrid services across computer rooms in two separate buildings (FCC-2 and GCC-B):

- This project was completed on 7-Jun-2011 (and tested by a building failure less than two hours later),
- The FermiGrid-HA2 infrastructure worked exactly as designed.

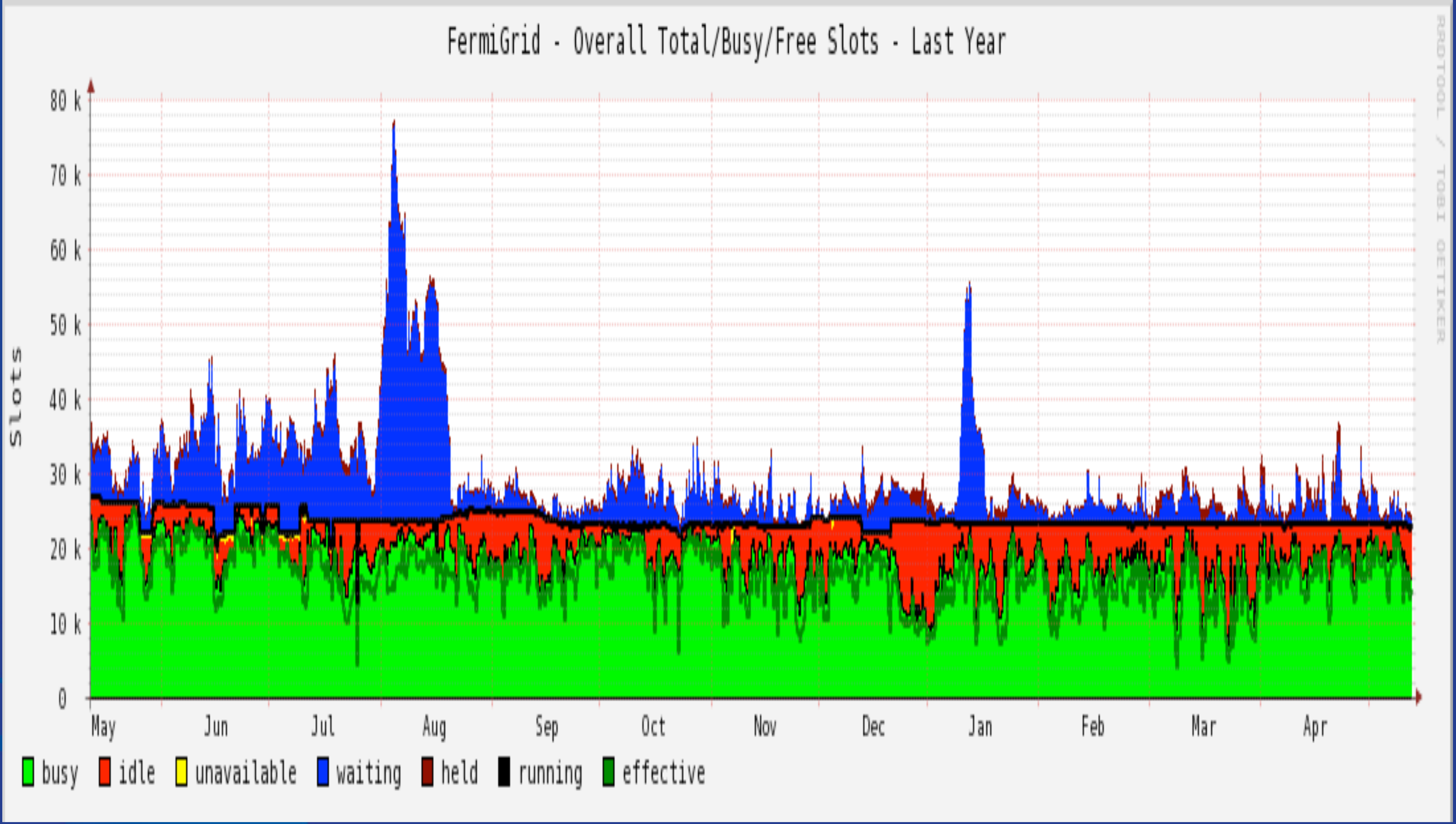
Our operational experience with FermiGrid-[HA,HA2] has shown the benefits of virtualization and service redundancy:

- Benefits to the user community – increased service reliability and uptime,
- Benefits to the service maintainers – flexible scheduling of maintenance and upgrade activities.

FermiGrid Service Availability (measured over the past year)

Service	Raw Availability	HA Configuration	Measured HA Availability	Minutes of Downtime
VOMS – VO Management Service	99.740%	Active-Active	99.988%	60
GUMS – Grid User Mapping Service	99.863%	Active-Active	100.000%	0
SAZ – Site AuthoriZation Service	99.864%	Active-Active	100.000%	0
Squid – Web Cache	99.817%	Active-Active	99.988%	60
MyProxy – Grid Proxy Service	99.751%	Active-Standby	99.874%	660
ReSS – Resource Selection Service	99.915%	Active-Active	99.988%	60
Gratia – Fermilab and OSG Accounting	99.662%	Active-Standby	99.985%	300
MySQL Database	99.937%	Active-Active	100.000%	0

Fermilab Campus Grid Utilization



Fermilab Campus Grid Statistics (as of April 2013)

Cluster(s)	Batch System	Job Slots	Raw Occupancy	Effective Utilization
CDF (Merged)	Condor	5268	81.8	70.0
CMS T1	Condor	6,272	90.3	85.5
D0 (Merged)	PBS	5,920	74.2	53.7
GP Grid	Condor	5,474	77.3	61.8
__________		__________	__________	__________
Overall-Today		22,934	80.4	68.3

FermiCloud

Infrastructure as a Service
Cloud Computing

Fermilab

Experience with FermiGrid = Drivers for FermiCloud

Access to pools of resources using common interfaces:

- Monitoring, quotas, allocations, accounting, etc.

Opportunistic access:

- Users can use the common interfaces to "burst" to additional resources to meet their needs

Efficient operations:

- Deploy common services centrally

High availability services:

- Flexible and resilient operations

Additional Drivers for FermiCloud

Existing development and integration facilities were:

- Technically obsolescent and unable to be used effectively to test and deploy the current generations of Grid middleware,
- The hardware was over 8 years old and was falling apart,
- The needs of the developers and service administrators in the Grid and Cloud Computing Department for reliable and "at scale" development and integration facilities were growing,
- Operational experience with FermiGrid had demonstrated that virtualization could be used to deliver production class services.

Additional Developments 2004-2012

Large multi-core servers have evolved from from 2 to 64 cores per box,

- A single "rogue" (poorly coded) user/application can impact 63 other users/applications.
- Virtualization can provide mechanisms to securely isolate users/applications.

Typical "bare metal" hardware has significantly more performance than usually needed for a single-purpose server,

- Virtualization can provide mechanisms to harvest/utilize the remaining cycles.

Complicated software stacks are difficult to distribute on grid,

- Distribution of preconfigured virtual machines together with GlideinWMS can aid in addressing this problem.

Large demand for transient development/testing/integration work,

- Virtual machines are ideal for this work.

Science is increasingly turning to complex, multiphase workflows.

- Virtualization coupled with cloud can provide the ability to flexibly reconfigure hardware "on demand" to meet the changing needs of science.

FermiCloud – Hardware Acquisition

In the FY2009 budget input cycle (in Jun-Aug 2008), based on the experience with the "static" virtualization in FermiGrid, the (then) Grid Facilities Department proposed that hardware to form a "FermiCloud" infrastructure be purchased,

- This was ***not*** funded for FY2009.

Following an effort to educate the potential stakeholder communities during FY2009, the hardware request was repeated in the FY2010 budget input cycle (Jun-Aug 2009),

- This time the request was funded.

The hardware specifications were developed in conjunction with the GPCF acquisition, and the hardware was delivered in May 2010.

FermiCloud – Initial Project Specifications

Once funded, the FermiCloud Project was established with the goal of developing and establishing Scientific Cloud capabilities for the Fermilab Scientific Program,

- Building on the very successful FermiGrid program that supports the full Fermilab user community and makes significant contributions as members of the Open Science Grid Consortium.
- Reuse High Availabilty, AuthZ/AuthN, Virtualization from Grid

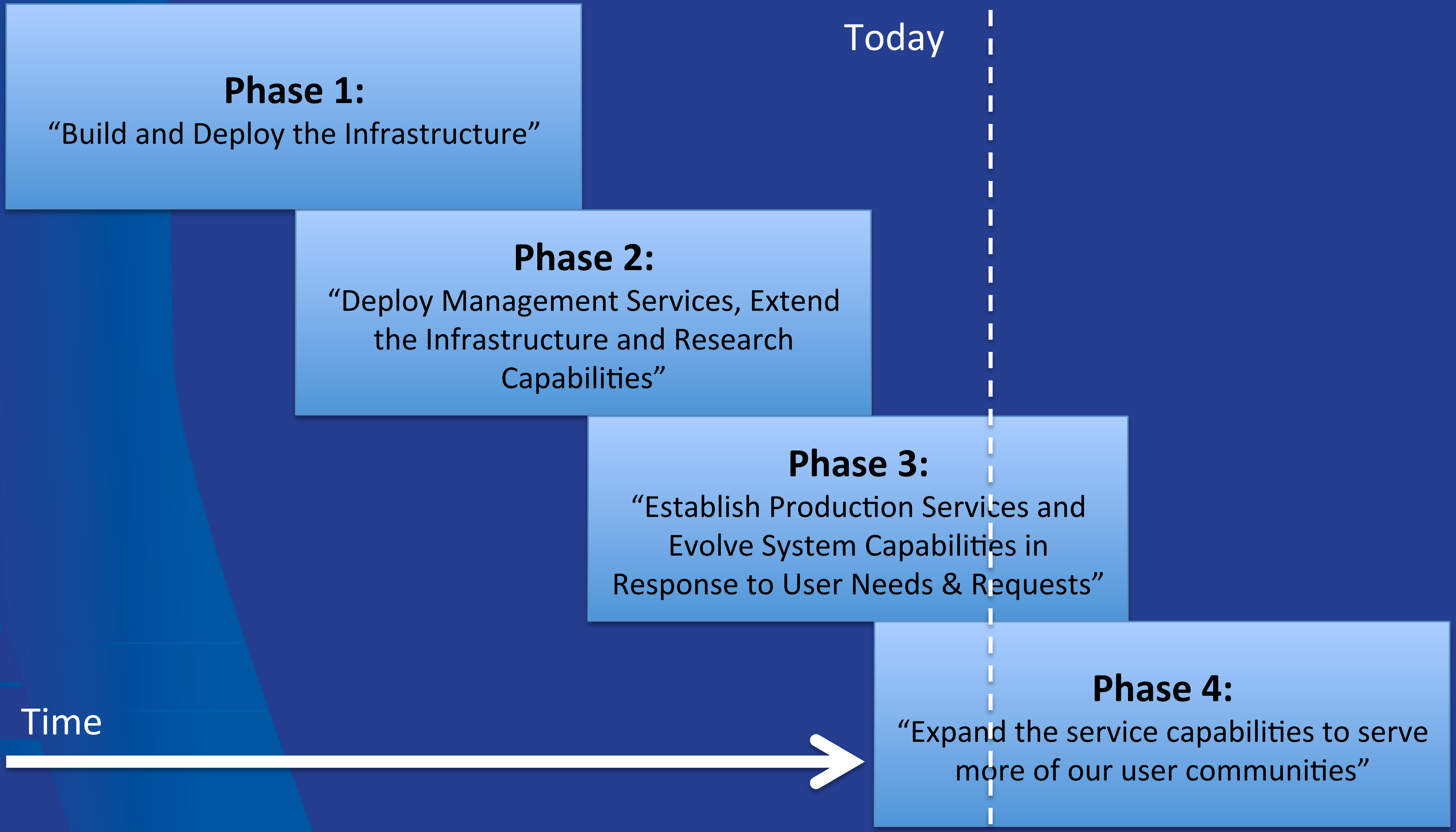
In a (very) broad brush, the mission of the FermiCloud project is:

- To deploy a production quality Infrastructure as a Service (IaaS) Cloud Computing capability in support of the Fermilab Scientific Program.
- To support additional IaaS, PaaS and SaaS Cloud Computing capabilities based on the FermiCloud infrastructure at Fermilab.

The FermiCloud project is a program of work that is split over several overlapping phases.

- Each phase builds on the capabilities delivered as part of the previous phases.

Overlapping Phases



Today

Phase 1:
"Build and Deploy the Infrastructure"

Phase 2:
"Deploy Management Services, Extend the Infrastructure and Research Capabilities"

Phase 3:
"Establish Production Services and Evolve System Capabilities in Response to User Needs & Requests"

Phase 4:
"Expand the service capabilities to serve more of our user communities"

Time

FermiCloud Phase 1: "Build and Deploy the Infrastructure"

- Specify, acquire and deploy the FermiCloud hardware,
- Establish initial FermiCloud requirements and selected the "best" open source cloud computing framework that met these requirements (OpenNebula),
- Deploy capabilities to meet the needs of the stakeholders (JDEM analysis development, Grid Developers and Integration test stands, Storage/dCache Developers, LQCD testbed).

Completed late 2010

FermiCloud Phase 2: "Deploy Management Services, Extend the Infrastructure and Research Capabilities"

- Implement x509 based authentication (patches contributed back to OpenNebula project and are generally available in OpenNebula V3.2),
- Perform secure contextualization of virtual machines at launch,
- Perform virtualized filesystem I/O measurements,
- Develop (draft) economic model,
- Implement monitoring and accounting,
- Collaborate with KISTI personnel to demonstrate Grid and Cloud Bursting capabilities,
- Perform initial benchmarks of Virtualized MPI,
- Target "small" low-cpu-load servers such as Grid gatekeepers, forwarding nodes, small databases, monitoring, etc.,
- Begin the hardware deployment of a distributed SAN,
- Investigate automated provisioning mechanisms (puppet & cobbler).

Completed late 2012

FermiCloud Phase 3: "Establish Production Services and Evolve System Capabilities in Response to User Needs & Requests"

- Deploy highly available 24x7 production services,
  - Both infrastructure and user services (Done).
- Deploy puppet & cobbler in production,
  - Done.
- Develop and deploy real idle machine detection,
  - Idle VM detection tool written by summer student.
- Research possibilities for a true multi-user filesystem on top of a distributed & replicated SAN,
  - GFS2 on FibreChannel SAN across FCC3 and GCC-B (Done).

Underway

- Live migration becomes important for this phase.
  - Manual migration has been used, Live migration is currently in test, Automatically triggered live migration yet to come.
- Formal ITIL Change Management "Go-Live",
  - Have been operating under "almost" ITIL Change Management for the past several months.

FermiCloud Phase 4: "Expand the service capabilities to serve more of our user communities"

- Complete the deployment of the true multi-user filesystem on top of a distributed & replicated SAN (Done),

- Demonstrate interoperability and federation:
  - Accepting VM's as batch jobs,
  - Interoperation with other Fermilab virtualization infrastructures (GPCF, VMware),
  - Interoperation with KISTI cloud, Nimbus, Amazon EC2, other community and commercial clouds,

- Participate in Fermilab 100 Gb/s network testbed,
  - Have installed 10 Gbit/second cards, about to be connected to 10 Gb/s switch ports.

- Perform more "Virtualized MPI" benchmarks and run some real world scientific MPI codes,
  - The priority of this work will depend on finding a scientific stakeholder that is interested in this capability.

- Reevaluate available open source Cloud computing stacks,
  - Including OpenStack,
  - We will also reevaluate the latest versions of Eucalyptus, Nimbus and OpenNebula.

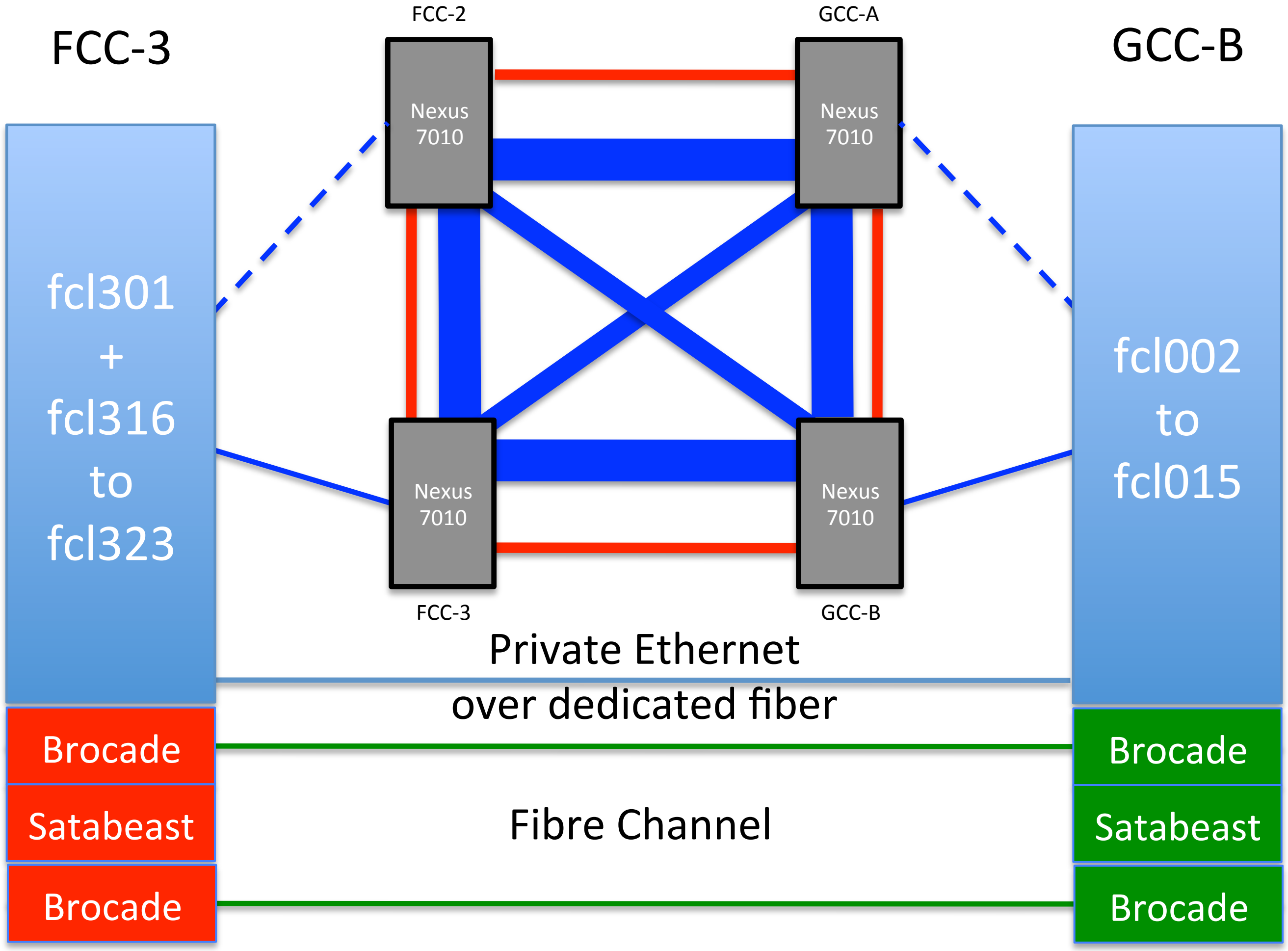
Specifications Nearly Final,
Work has Started

FermiCloud Phase 5

- See the program of work in the (draft) Fermilab-KISTI FermiCloud CRADA,
- This phase will also incorporate any work or changes that arise out of the Scientific Computing Division strategy on Clouds and Virtualization,
- Additional requests and specifications are also being gathered!

Specifications
Under Development

FermiCloud Network & SAN Topology



FY2011 / FY2012

Distributed Shared File System Benefits

Fast Launch:

- Almost immediate as compared to 3-4 minutes with ssh/scp,

Live Migration:

- Can move virtual machines from one host to another for scheduled maintenance, transparent to users,

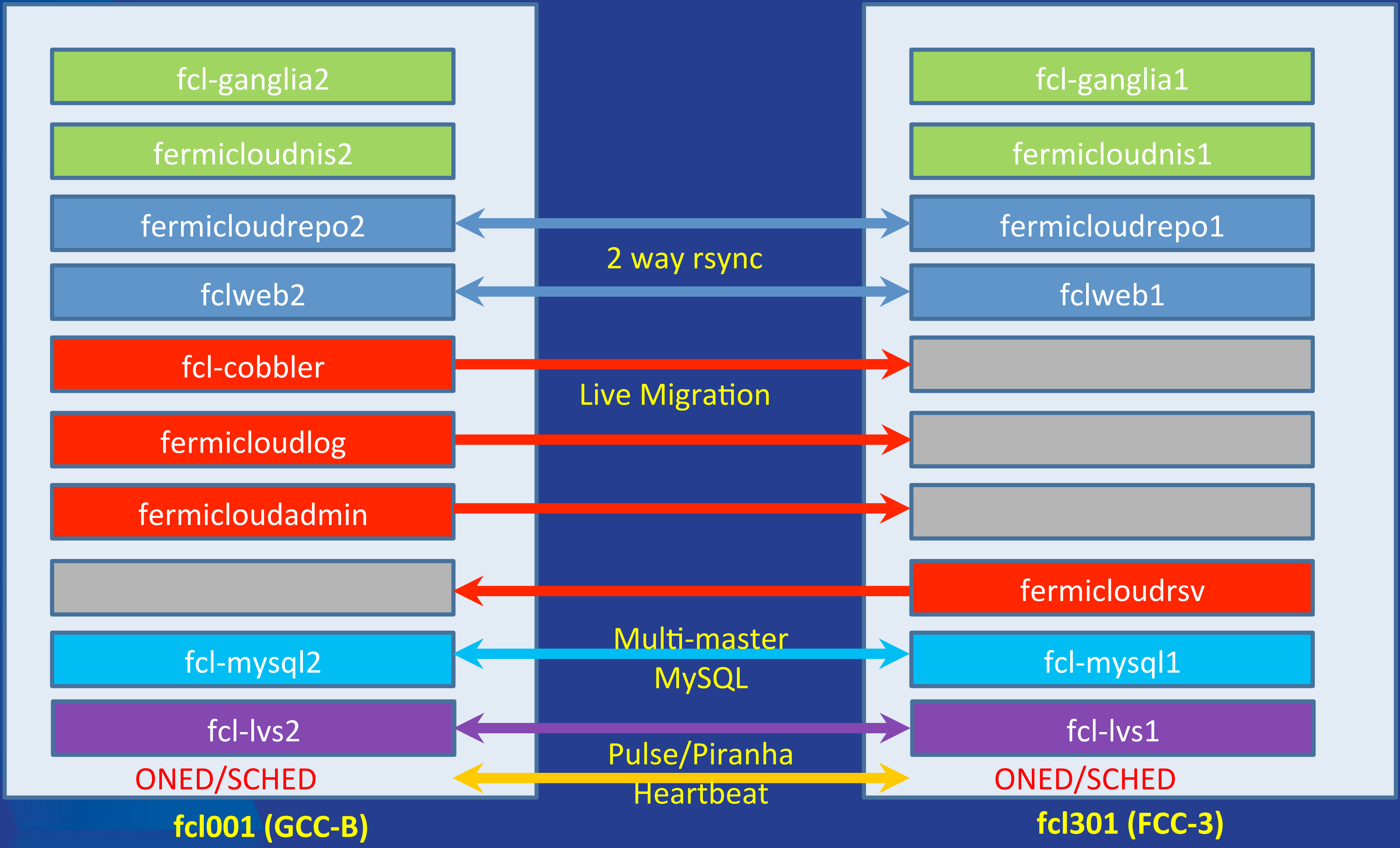
Persistent data volumes:

- Can move quickly with machines,

With mirrored volumes across SAN:

- Can relaunch virtual machines in surviving building in case of building failure/outage.

FermiCloud-HA
Head Node Configuration



Virtualized MPI Measurements (Note 1)

Configuration	#Host Systems	#VM/host	#CPU	Total Physical CPU	HPL Benchmark (Gflops)	Gflops/Core
Bare Metal without pinning	2	--	8	16	13.9	0.87
Bare Metal with pinning (Note 2)	2	--	8	16	24.5	1.53
VM without pinning (Notes 2,3)	2	8	1 vCPU	16	8.2	0.51
VM with pinning (Notes 2,3)	2	8	1 vCPU	16	17.5	1.09
VM+SRIOV with pinning (Notes 2,4)	2	7	2 vCPU	14	23.6	1.69

Notes: (1) Work performed by Dr. Hyunwoo Kim of KISTI in collaboration with Dr. Steven Timm of Fermilab.
(2) Process/Virtual Machine "pinned" to CPU and associated NUMA memory via use of numactl.
(3) Software Bridged Virtual Network using IP over IB (seen by Virtual Machine as a virtual Ethernet).
(4) SRIOV driver presents native InfiniBand to virtual machine(s), 2nd virtual CPU is required to start SRIOV, but is only a virtual CPU, not an actual physical CPU.

Current FermiCloud Capabilities

Public network access via the high performance Fermilab network,
- This is a distributed, redundant network.

Private 1 Gb/sec network,
- This network is bridged across FCC and GCC on private fiber,

High performance Infiniband network to support HPC based calculations,
- Performance under virtualization is ~equivalent to "bare metal",
- Currently split into two segments,
- Segments could be bridged via Mellanox MetroX.

Access to a high performance redundant FibreChannel based SAN,
- This SAN spans both buildings,

Access to the high performance BlueArc based filesystems,
- The BlueArc is located on FCC-2,

Access to the Fermilab dCache and enStore services,
- These services are split across FCC and GCC,

Access to 100 Gbit Ethernet test bed in LCC (via FermiCloud integration nodes),
- Intel 10 Gbit Ethernet converged network adapter X540-T1.

We have a (draft) Economic Model,
- Costs are competitive with commercial cloud providers,
- High performance network access to Fermilab data sets (without commercial cloud provider data movement charges).

FermiCloud Economic Model

Calculate rack cost:

- Rack, public Ethernet switch, private Ethernet switch, Infiniband switch,
- $11,000 USD (one time).

Calculate system cost:

- Based on 4 year lifecycle,
- $6,500 USD / 16 processors / 4 years => $125 USD / year

Calculate storage cost:

- 4 x FibreChannel switch, 2 x SATAbeast, 5 year lifecycle,
- $130K USD / 60 Tbytes / 5 years => $430 USD / TB-year

Calculate fully burdened system administrator cost:

- Current estimate is 400 systems per administrator,
- $250K USD / year / 400 systems => $750 USD / system-year

FermiCloud Draft Economic Model Results (USD)

SLA	24x7	9x5	Opportunistic
“Unit” (HT CPU + 2 GB)	$125	$45	$25
Add’l core	$125	$125	$125
Add’l memory per GB	$30	$30	$30
Add’l local disk per TB	$40	$40	$40
SAN disk per TB	$475	$475	$475
BlueArc per TB	$430	$430	$430
System Administrator	$750	$750	$750
Specialized Service Support	“Market”	“Market”	“Market”

Note - Costs in the above chart are per year

Korea Institute of Science and Technology Information (KISTI)

KISTI personnel have been collaborating with FermiCloud since the summer of 2011 on a variety of cloud computing developments:

- Grid and Cloud Bursting (vCluster),
- Virtualized MPI Measurements using Infiniband SRIOV drivers.

DOE has just approved a Cooperative Research and Development Agreement (CRADA) between Fermilab and KISTI:

- This CRADA defines the project, including the deliverables, scope and schedule.
- It also governs the relationship between FNAL and KISTI, including budget and intellectual property, and also addresses any potential environmental impacts.

The collaboration with KISTI has leveraged the resources and expertise of both institutions to achieve significant benefits

Virtual Machines as Jobs

OpenNebula (and all other open-source IaaS stacks) provide an emulation of Amazon EC2.

HTCondor team has added code to their "Amazon EC2" universe to support the X.509-authenticated protocol.

Planned use case for GlideinWMS to run Monte Carlo on clouds public and private.

Feature already exists,

- this is a testing/integration task only.

Grid Bursting

Seo-Young Noh, KISTI visitor @ FNAL, showed proof-of-principle of "vCluster" in summer 2011:

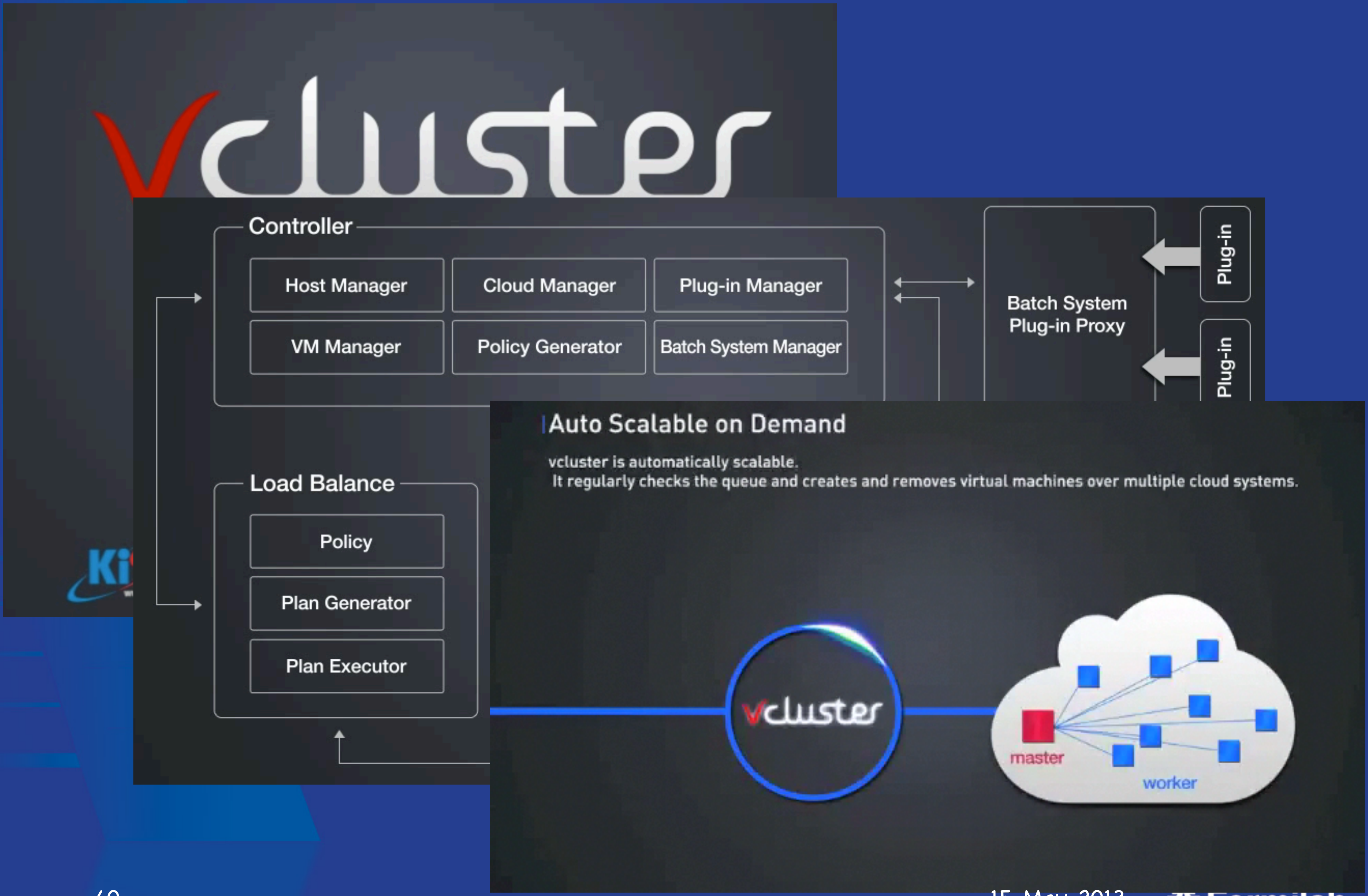
- Look ahead at Condor batch queue,
- Submit worker node virtual machines of various VO's to FermiCloud or Amazon EC2 based on user demand,
- Machines join grid cluster and run grid jobs from the matching VO.

Need to strengthen proof-of-principle, then make cloud slots available to FermiGrid.

Several other institutions have expressed interest in extending vCluster to other batch systems such as Grid Engine.

KISTI staff have a program of work for the development of vCluster.

vCluster at SC2012



vcluster

Controller

Host Manager

Cloud Manager

Plug-in Manager

VM Manager

Policy Generator

Batch System Manager

Batch System Plug-in Proxy

Plug-in

Plug-in

Load Balance

Policy

Plan Generator

Plan Executor

Auto Scalable on Demand

vcluster is automatically scalable.
It regularly checks the queue and creates and removes virtual machines over multiple cloud systems.

Cloud Bursting

OpenNebula already has built-in "Cloud Bursting" feature to send machines to Amazon EC2 if the OpenNebula private cloud is full.

Will evaluate/test it as part of the KISTI CRADA,

- To see if it meets our technical and business requirements, or if something else is necessary,
- Also will test interoperability against other cloud computing stacks (OpenStack, NIMBUS, Eucalyptus).

True Idle VM Detection

In times of resource need, we want the ability to suspend or "shelve" idle VMs in order to free up resources for higher priority usage.

- This is especially important in the event of constrained resources (e.g. during building or network failure).

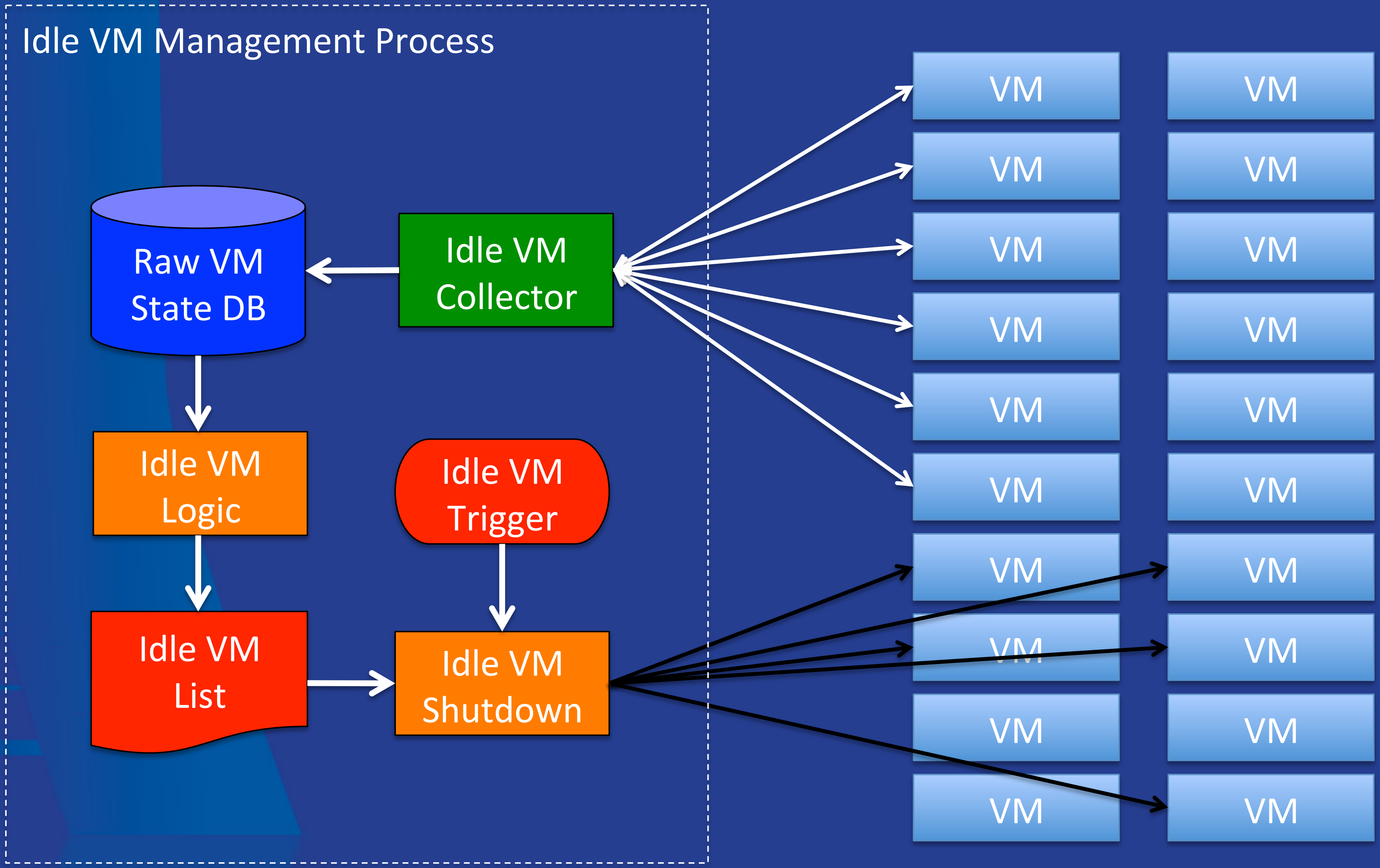
Shelving of "9x5" and "opportunistic" VMs allows us to use FermiCloud resources for Grid worker node VMs during nights and weekends

- This is part of the draft economic model.

During the summer of 2012, an Italian co-op student wrote (extensible) code for an "Idle VM Probe" that can be used to detect idle virtual machines based on CPU, disk I/O and network I/O.

- This is the biggest pure coding task left in the FermiCloud project,
- Work will be performed by a consultant funded from the Fermilab-KISTI CRADA.

Idle VM Information Flow



Interoperability and Federation

Driver:

- Global scientific collaborations such as LHC experiments will have to interoperate across facilities with heterogeneous cloud infrastructure.

European efforts:

- EGI Cloud Federation Task Force – several institutional clouds (OpenNebula, OpenStack, StratusLab).
- HelixNebula—Federation of commercial cloud providers

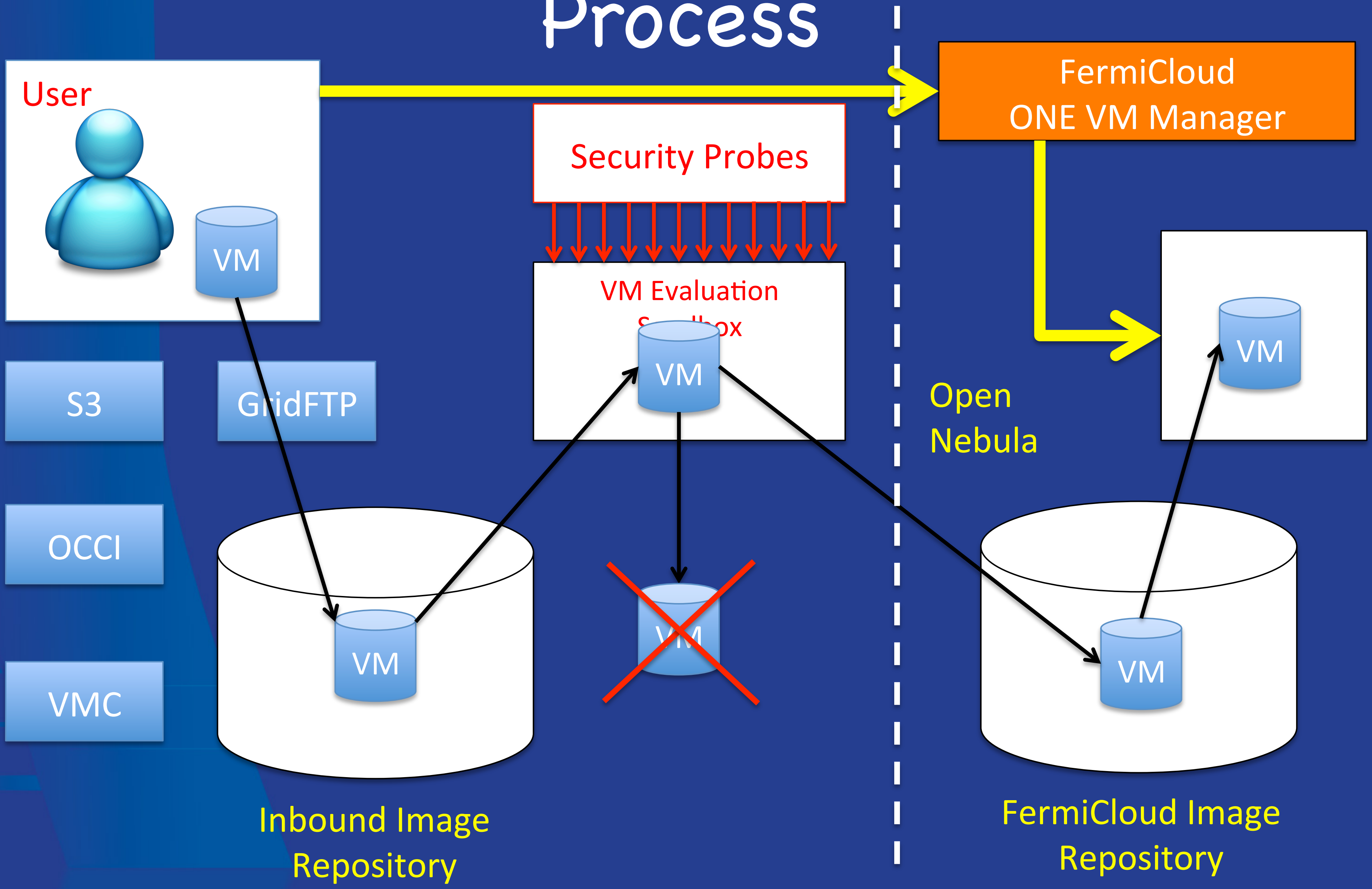
Our goals:

- Show proof of principle—Federation including FermiCloud + KISTI "G Cloud" + one or more commercial cloud providers + other research institution community clouds if possible.
- Participate in existing federations if possible.

Core Competency:

- FermiCloud project can contribute to these cloud federations given our expertise in X.509 Authentication and Authorization, and our long experience in grid federation

Possible VM Image Acceptance Process



High-Throughput Fabric Virtualization

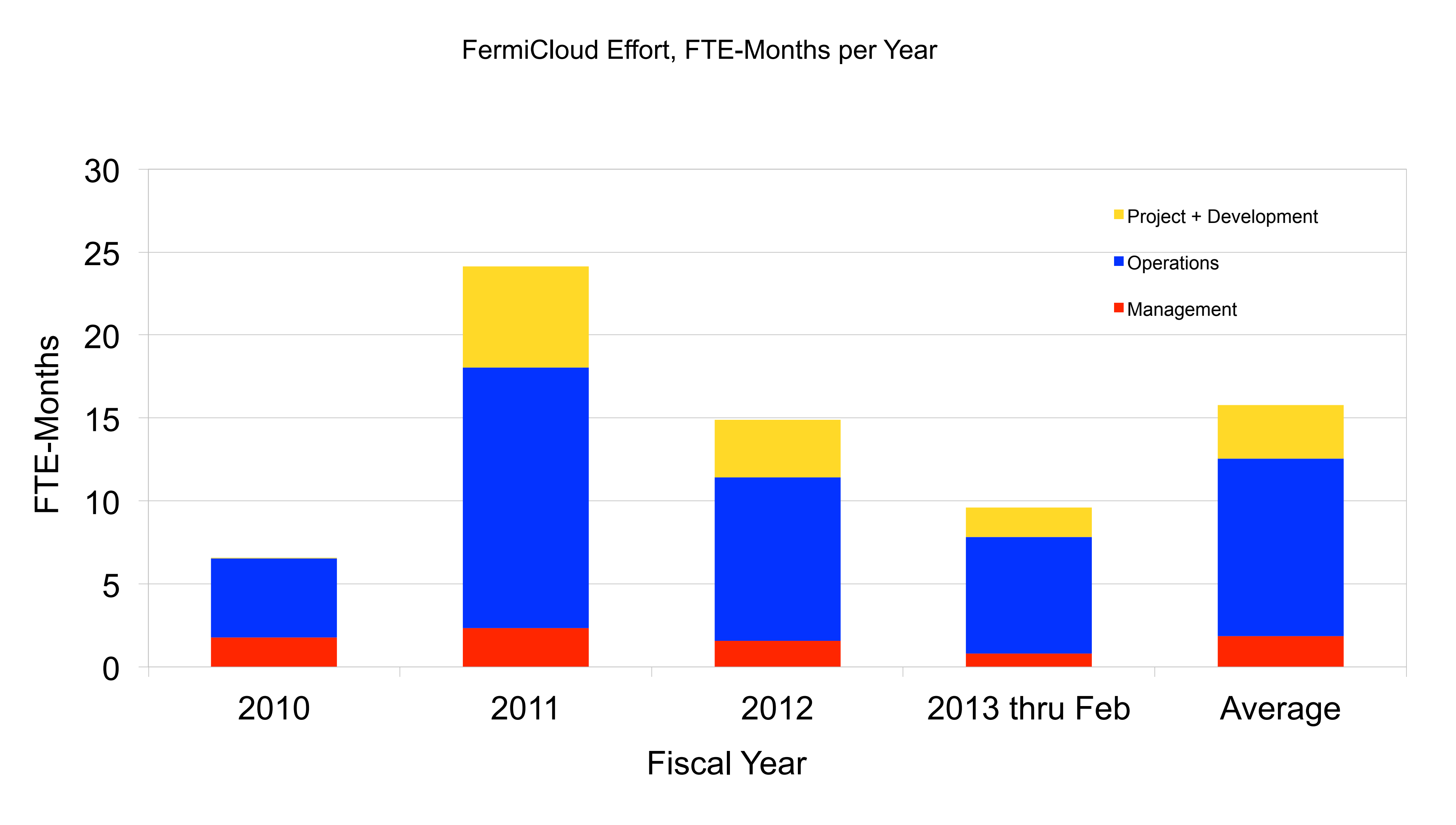
Follow up earlier virtualized MPI work:

- Use it in real scientific workflows,
- Example – simulation of data acquisition systems (the existing FermiCloud Infiniband fabric has already been used for such).

Will also use FermiCloud machines on 100Gbit Ethernet test bed

- Evaluate / optimize virtualization of 10G NIC for the use case of HEP data management applications,
- Compare and contrast against Infiniband.

FermiCloud FTE Effort Plot



Current FermiCloud Stakeholders

Grid & Cloud Computing Personnel,

Run II – CDF & D0,

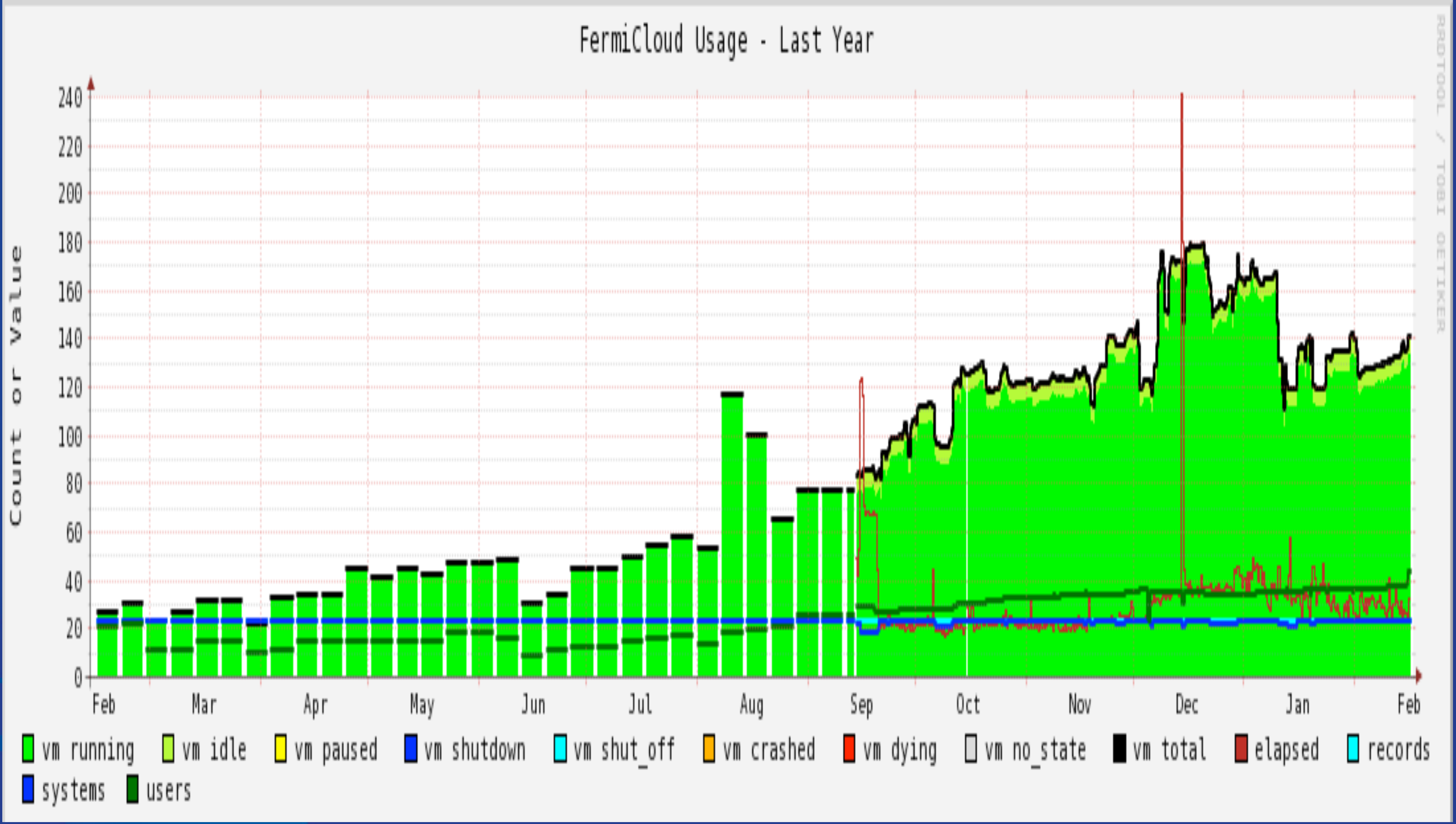
Intensity Frontier Experiments,

Cosmic Frontier (LSST),

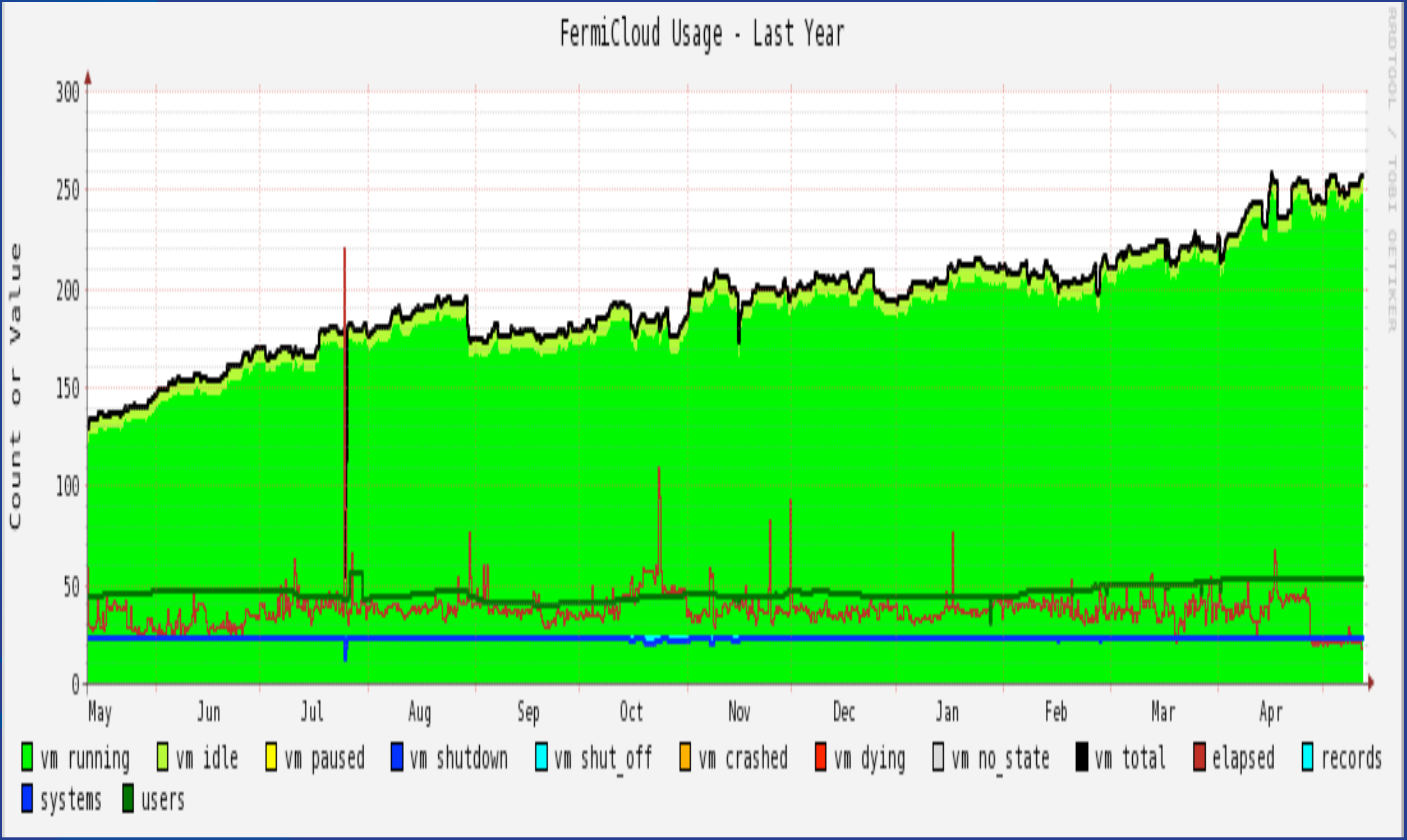
Korea Institute of Science & Technology Investigation (KISTI),

Open Science Grid (OSG) software refactoring from pacman to RPM based distribution, Grid middleware development.

Initial FermiCloud Utilization



Current FermiCloud Utilization



General Physics Computing Facility (GPCF)

GPCF Service Providers

Running Experiments (REX) Department

Fermilab Experimental Facilities (FEF) Department

GPCF Stakeholders

Intensity Frontier experiments

- ArgoNeuT, Muon g-2, Liquid argon R&D, LBNE, MicroBooNE, MiniBooNE, MINERvA, MINOS, Mu2e, NOvA, SeaQuest

Cosmic Frontier experiments

- Auger, COUPP, DarkSide-50, LSST

Energy Frontier (Run II) experiments

- CDF & D0

Typical GPCF Applications

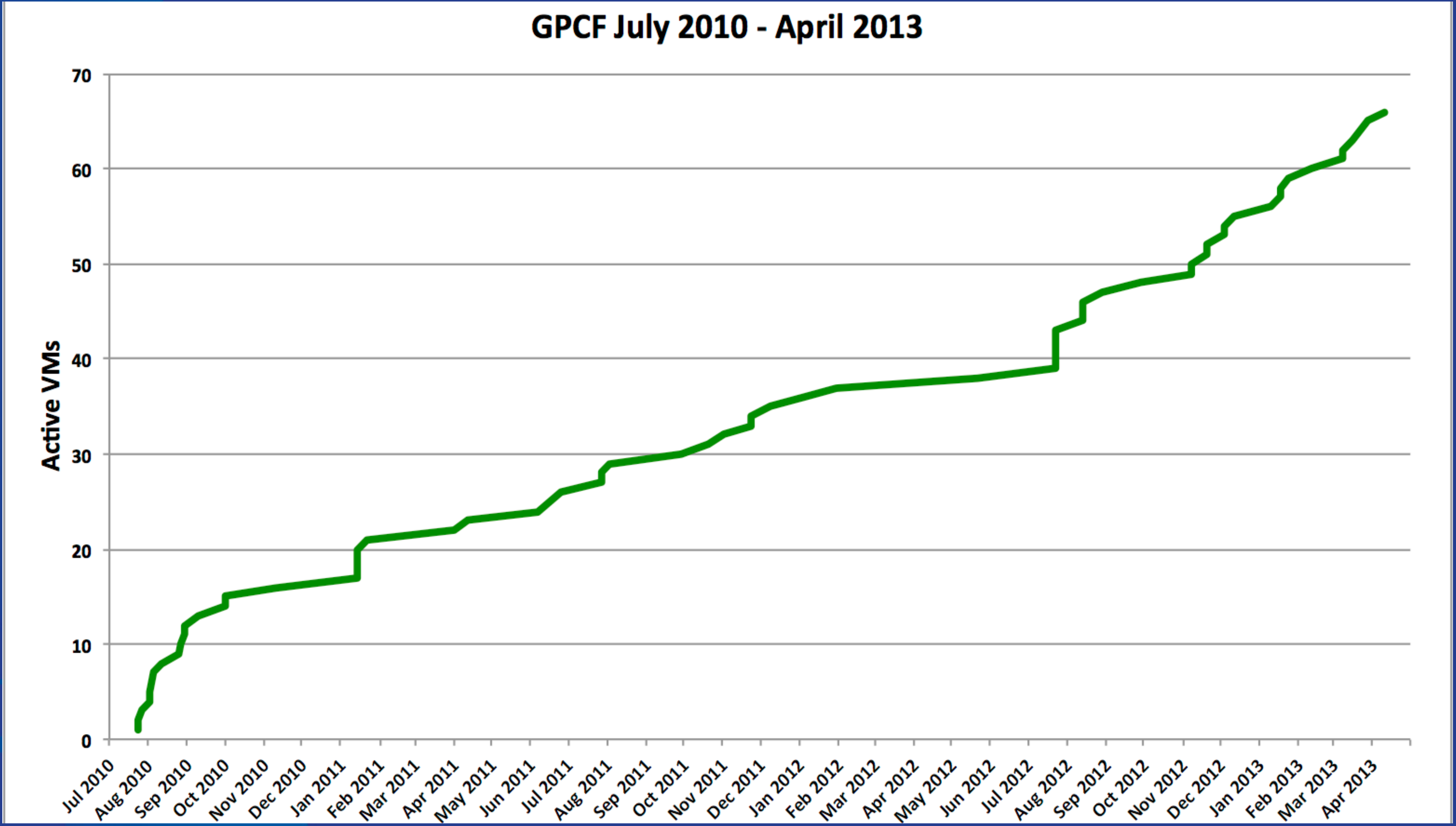
Interactive sessions

Analysis software: ROOT, experiment-specific code, etc.

Web servers

Batch/workload management: HTCondor

GPCF Utilization



VMware Virtual Services

Virtual Services Stakeholders

Finance Section,

Business Services Section,

Directorate (Budget Office, Internal Audit, Legal, Visual Media Services)

Technical Division,

FESS,

ES&H,

Computer Security,

Authentication Services,

Network Services,

Web & Collaboration,

Desktop Engineering

Virtual Services Applications

Redmine, GIT, SVN (code management)

Sharepoint cluster (DEV,INT,PRD instances)

PeachTree Accounting

Filemaker

Windows and Linux Print Servers

DNS Guest Registration

Teamcenter (Windows and Linux)

FIDM (Fermilab Indentity Management)

Peoplesoft (HRMS)

Plone

DOORS (Requirements Management Software)

Teammate (Enterprise Audit Software)

Indico

Jabber

VoIP

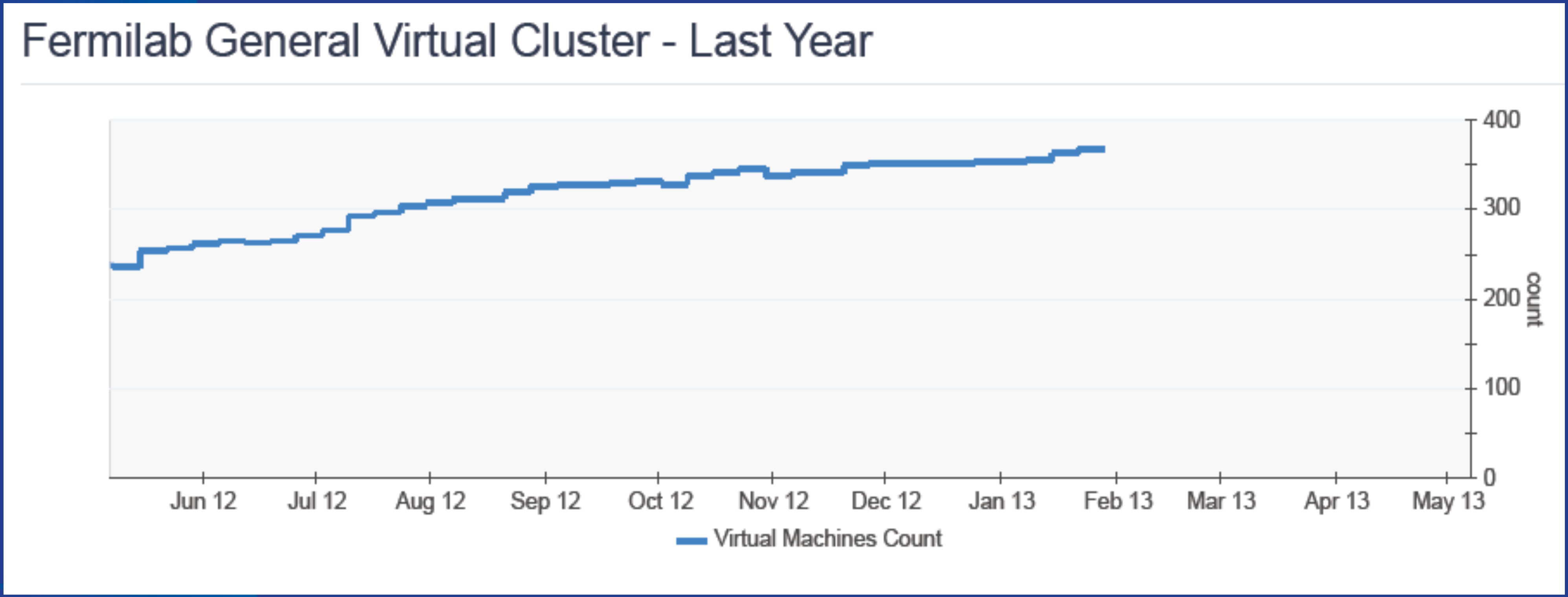
Tissue, Nodelocator, MRTG, NGOP

Fermi Time & Labor Terminal Server cluster

Promise Servers

20+ Windows Terminal Servers

40-60 virtual desktops (Windows 7)



Fermilab General Virtual Cluster - Last Year

Plans for the Future

Significant Benefits from Virtualization & Cloud Computing

- Grid & Cloud Computing middleware development,
- OSG Software Development,
- X.509 AuthZ/AuthN contributions to OpenNebula,
- NFS V4.1 Tests with dCache (Mike Wang),
- Forthcoming 100 Gbit testing,
- Additional benchmarking of virtualized MPI using Infiniband,
- Modeling DAQ systems using Infiniband,
- Benchmarking virtualized storage,
- "At scale" Grid and cloud middleware testing,
- Operation of production services across outages,
- Economic model shows Raw VM costs are comparable to Amazon EC2, with much better access to Fermilab data sets (and we don't have to pay the commercial provider I/O charges),
- Highly available & resilient ***production*** Cloud infrastructure.

Lessons - Nov 2004 to Jan 2013

Drivers that were pushing the adoption of Grid in 2004/2005 are drivers that are pushing the adoption of Cloud today,

FermiGrid and FermiCloud projects have demonstrated significant cross pollination:

- Knowledge learned from each project serves to enable progress on the other project,
- Total operations effort (the sum of FermiGrid + FermiCloud) has remained constant (modulo personnel departures).

Evolution of Provisioning - 1

We are at a transition point in how systems and services can be (and are) provisioned.

Traditional "static" hardware provisioning:

- Specify and buy a system, rack it, commission it, run it until retirement,
- Repeat, repeat, repeat…

"Static" virtual machine provisioning:

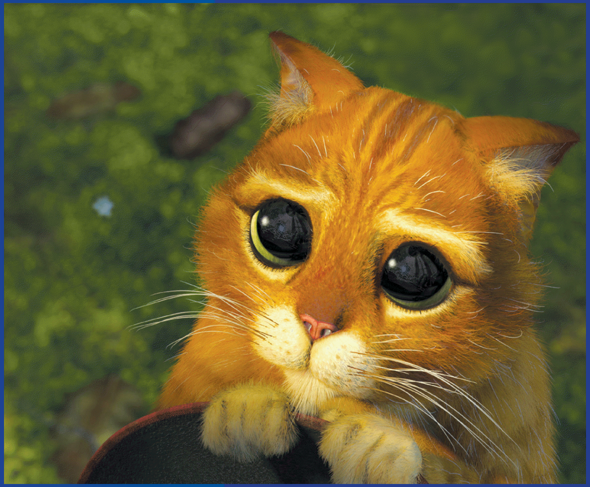
- FermiGrid, GPCF and Vmware Virtual Services are examples,
- All offer significant cost and flexibility advantages over static hardware provisioning,
- Benchmarks run by GCC personnel have shown that virtual machine performance is equivalent to the "bare metal" performance (CPU, Network, Disk, Database, HPC/MPI).

Evolution of Provisioning - 2

Cloud based virtual machine provisioning:

- Virtual machine provisioning on demand (no wait for procurement cycle),
- Cloud infrastructure also supports "static" (long term) provisioning,
- Built in virtual machine lifecycle,
- Built in support for high availability and service recovery,
- Allows the scientific workflows to flexibly reconfigure the infrastructure to match the scientific needs without the need to go move racks, pull cables and re-image the server, just by loading different virtual machines, all in a common environment,
- Use a common set of management tools and centrally managed services,
- Securely run older (out of support) operating systems on virtual machines,
- Can be evolved incrementally as needs and requirements develop,
- Can support heterogeneous hardware capabilities,
- All performed using a common environment for the users and service providers!

CERN Agile Infrastructure Service Model – Tim Bell



Pets:

- Pets are given names like pussinboots.cern.ch
- They are unique, lovingly hand raised and cared for
- When they get ill, you nurse them back to health

Cattle:

- Cattle are given numbers like vm0042.cern.ch
- They are almost identical to other cattle
- When they get ill, you get another one

Future application architectures tend towards Cattle but Pet support is needed for some specific zones of the cloud

Summary

Virtualization & Cloud Computing

Virtualization is a significant component of our core computing (business) and scientific computing continuity planning,

Virtualization allows us to consolidate multiple functions into a single physical server, saving power & cooling,

Physical to virtual (P2V), virtual to virtual (V2V) and virtual to physical (V2P) migrations are supported,

It's significantly easier and faster (and safer too...) to "lift" virtual machines rather than physical machines.

Benefits to Science

Science is directly and indirectly benefiting from the virtualization and cloud computing infrastructure (FermiGrid, FermiCloud, GPCF, Virtual Services) at Fermilab:

- Energy Frontier - CDF, D0, CMS, Atlas, etc.,
- Intensity Frontier – Minos, MiniBoone, MicroBoone, NOvA, LBNE, etc.,
- Cosmic Frontier – DES, LSST, etc.,
- Open Science Grid, KISTI, etc.

FermiCloud Summary

FermiCloud operates at the forefront of delivering cloud computing capabilities to support scientific research:

- By starting small, developing a list of requirements, building on existing Grid knowledge and infrastructure to address those requirements, FermiCloud has managed to deliver a production class Infrastructure as a Service cloud computing capability that supports science at Fermilab.
- FermiCloud has provided FermiGrid with an infrastructure that has allowed us to test Grid middleware at production scale prior to deployment.
- The Open Science Grid software team used FermiCloud resources to support their RPM "refactoring" and is currently using it to support their ongoing middleware development/integration.

GPCF Summary

GPCF provides a flexible facility to meet the central computing needs of experiments at Fermilab:

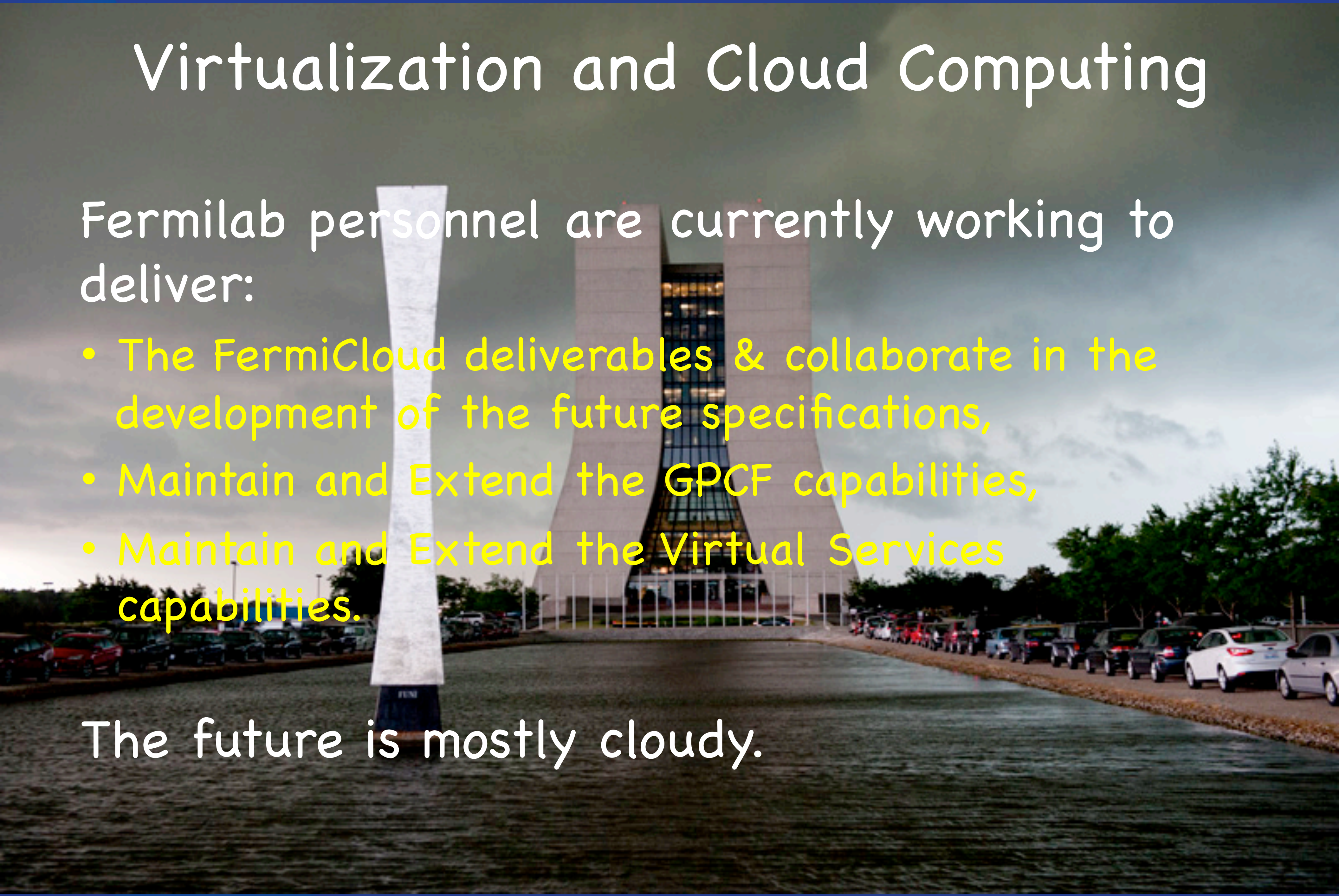
- GPCF uses tools included in Scientific Linux (KVM, libvirt, Clustered LVM) to implement an environment that allows for rapid, easily configurable provisioning of servers to meet experiment's needs.
- GPCF virtual servers are intended for medium to long-term use and are supported by the same tools and the same system administration group as traditional, physical servers.
- Intensity Frontier and other experiments use GPCF nodes for interactive computing - event display, job submission, collaboration, software development, etc. - and other uses.
- GPCF also provides platforms for monitoring and other service applications.

Virtual Services Summary

The VMware Virtual Services installation provides a flexible facility to support the "core" (business & management) computing of Fermilab.

- Virtual Services use VMware and associated tools in order to operate in a configuration that 3rd party vendors (Oracle, Peoplesoft, etc.) are willing to support.

Virtualization and Cloud Computing



Fermilab personnel are currently working to deliver:

- The FermiCloud deliverables & collaborate in the development of the future specifications,
- Maintain and Extend the GPCF capabilities,
- Maintain and Extend the Virtual Services capabilities.

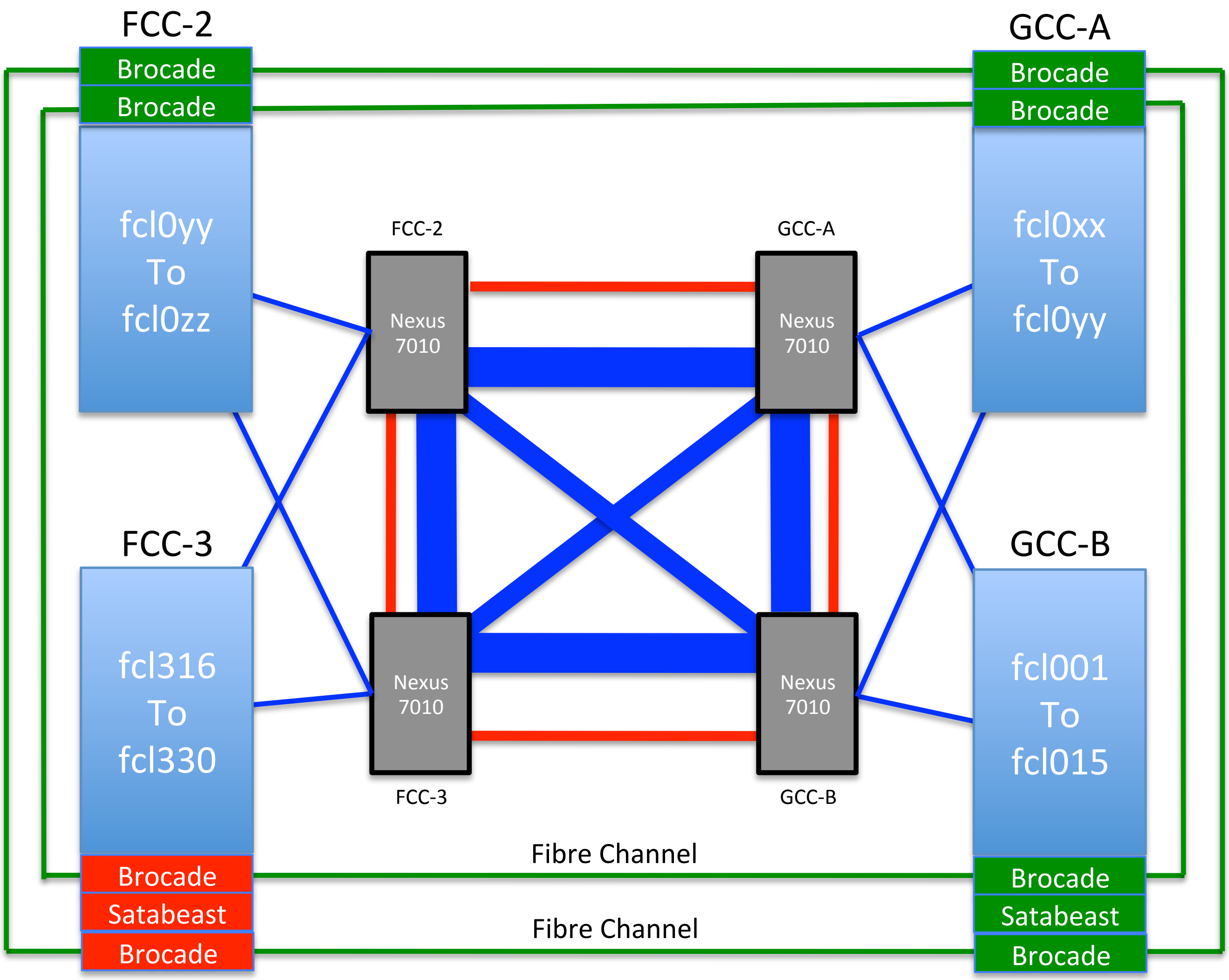
The future is mostly cloudy.

Thank You

Any Questions?

Extra Slides

FermiCloud – Network & SAN (Possible Future – FY2013/2014)



FermiCloud and HPC

In 2010, the HPC Department signed on as one of the initial stakeholders of FermiCloud.

The goal was to put together a basic "training/learning" environment on FermiCloud to allow new HPC researchers to make their mistakes on FermiCloud rather than impacting the production HPC resources.

FermiCloud – Infiniband

To enable HPC, the FermiCloud hardware specifications included the Mellanox SysConnect II Infiniband card that was claimed by Mellanox to support virtualization with the Infiniband SRIOV driver.

Unfortunately, despite promises from Mellanox, we were unable to take delivery of the Infiniband SRIOV driver while working through the standard sales support channels.

While at SuperComputing 2011 in Seattle in November 2011, Steve Timm, Amitoj Singh and I met with the Mellanox engineers and were able to make arrangements to receive the Infiniband SRIOV driver.

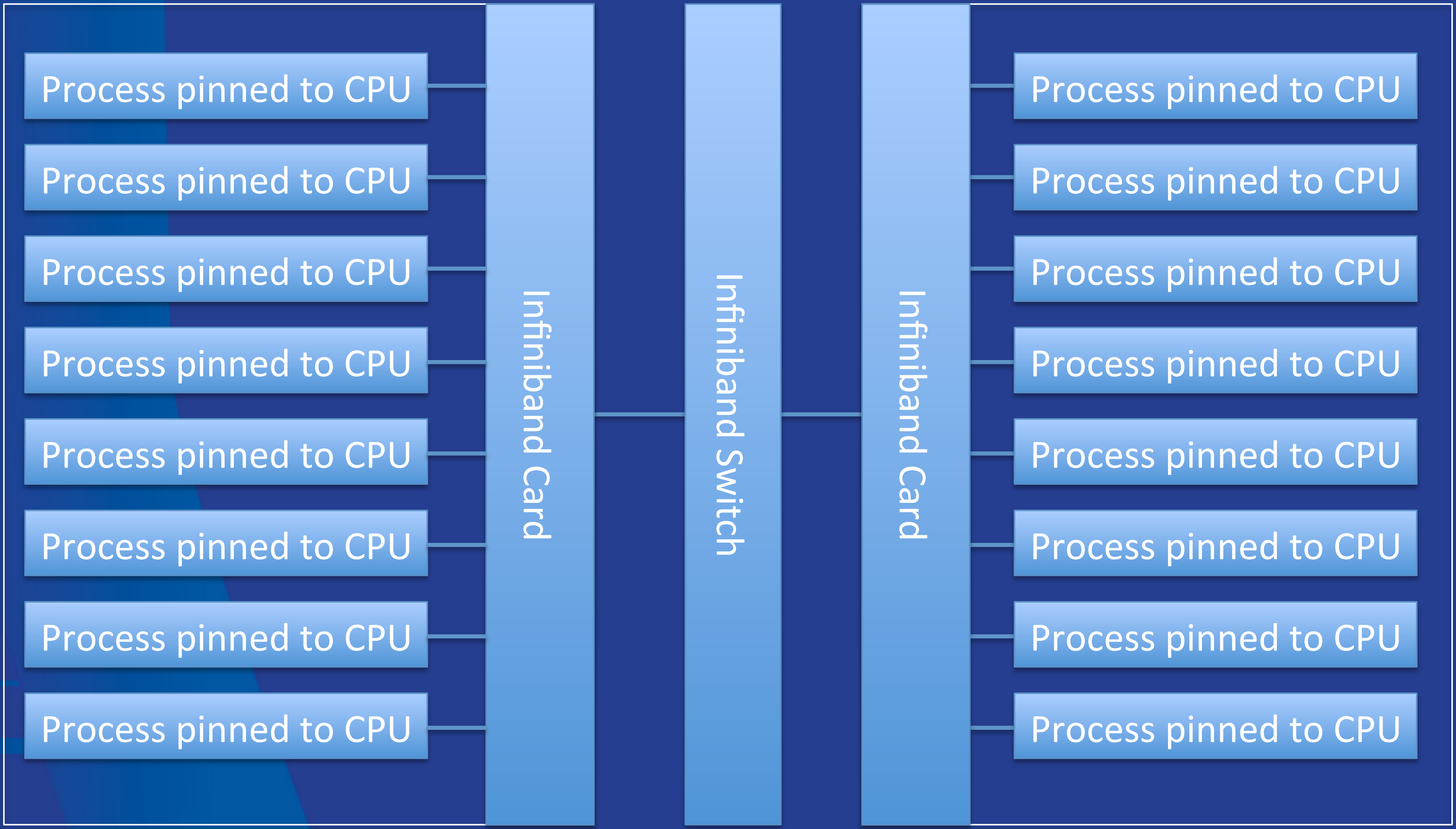
The Infiniband SRIOV driver was delivered to us in December 2011.

FermiCloud – SRIOV Driver

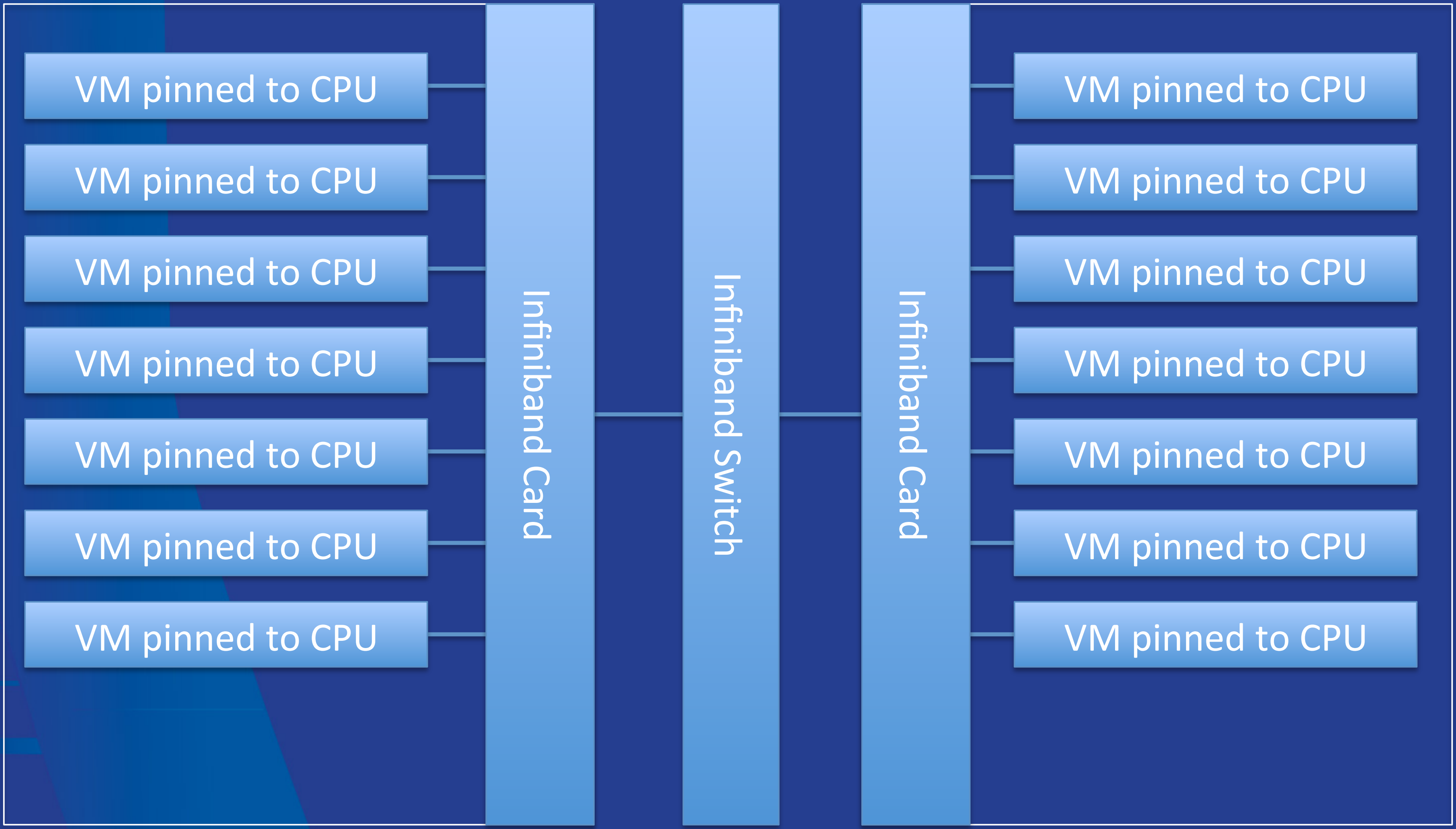
Using the Infiniband SRIOV driver, we were able to make measurements comparing MPI on “bare metal” to MPI on KVM virtual machines on the identical hardware.

This allows a direct measurement of the MPI “virtualization overhead”.

"Bare Metal"


Process pinned to CPU
Process pinned to CPU
Process pinned to CPU
Process pinned to CPU
Process pinned to CPU
Process pinned to CPU
Process pinned to CPU
Process pinned to CPU
Infiniband Card
Infiniband Switch
Infiniband Card
Process pinned to CPU
Process pinned to CPU
Process pinned to CPU
Process pinned to CPU
Process pinned to CPU
Process pinned to CPU
Process pinned to CPU
Process pinned to CPU

“Virtual MPI”


VM pinned to CPU
VM pinned to CPU
VM pinned to CPU
VM pinned to CPU
VM pinned to CPU
VM pinned to CPU
VM pinned to CPU
Infiniband Card
Infiniband Switch
Infiniband Card
VM pinned to CPU
VM pinned to CPU
VM pinned to CPU
VM pinned to CPU
VM pinned to CPU
VM pinned to CPU
VM pinned to CPU

MPI on FermiCloud (Note 1)

Configuration	#Host Systems	#VM/host	#CPU	Total Physical CPU	HPL Benchmark (Gflops)	Gflops/Core
Bare Metal without pinning	2	--	8	16	13.9	0.87
Bare Metal with pinning (Note 2)	2	--	8	16	24.5	1.53
VM without pinning (Notes 2,3)	2	8	1 vCPU	16	8.2	0.51
VM with pinning (Notes 2,3)	2	8	1 vCPU	16	17.5	1.09
VM+SRIOV with pinning (Notes 2,4)	2	7	2 vCPU	14	23.6	1.69

Notes: (1) Work performed by Dr. Hyunwoo Kim of KISTI in collaboration with Dr. Steven Timm of Fermilab.
(2) Process/Virtual Machine “pinned” to CPU and associated NUMA memory via use of numactl.
(3) Software Bridged Virtual Network using IP over IB (seen by Virtual Machine as a virtual Ethernet).
(4) SRIOV driver presents native InfiniBand to virtual machine(s), 2nd virtual CPU is required to start SRIOV, but is only a virtual CPU, not an actual physical CPU.

Summary & Near Term Work

Using the Mellanox Infiniband SRIOV drivers on FermiCloud virtual machines, MPI performance with the HPL benchmark has been demonstrated to be >96% of the native "bare metal" performance!

- Note that this HPL benchmark performance measurement was accomplished using **2 fewer** physical CPUs than the corresponding "bare metal" performance measurement!
- The **2 fewer** physical CPUs is a limitation of the current Infiniband SRIOV driver, if we can get this fixed then the virtualization results are likely to be able to achieve full parity with the "bare metal" results.

Near Term Work – Extend the MPI measurements:

- 4 & 8 physical systems (32 & 64 cores),
- 4 & 8 vm host systems (28 & 56 cores).

Security Contributions

Security Policy

- Proposed Cloud Computing Environment

X.509 Authentication and Authorization

Secure Contextualization

Participation in HEPiX Virtualisation taskforce

Cloud Computing Environment

- FermiCloud Security taskforce recommended to CSBoard/CST that a new Cloud Computing Environment be established
- This is currently under preparation.
- Normal FermiCloud use is authenticated by Fermi Kerberos credentials, \
  - either X.509 or MIT Kerberos or both.
- Special concerns:
  - Users have root:
  - Usage can be a combination of
    - Grid usage (Open Science Environment) and
    - Interactive usage (General Computing Environment)
  - Planning for "secure cloud" to handle expected use cases:
    - Archival systems at old patch levels or legacy OS.
    - Data and code preservation systems.
    - Non-baselined OS (Ubuntu, Centos, SUSE)
    - Non-kerberos services which can live only on private net.

OpenNebula Authentication

OpenNebula came with "pluggable" authentication, but few plugins initially available. OpenNebula 2.0 Web services by default used access key / secret key mechanism similar to Amazon EC2. No https available.

Four ways to access OpenNebula

- Command line tools
- Sunstone Web GUI
- "ECONE" web service emulation of Amazon Restful (Query) API
- OCCI web service.

Fermilab wrote X.509-based authentication plugins.

- Patches to OpenNebula to support this were developed at Fermilab and submitted back to the OpenNebula project in Fall 2011 (generally available in OpenNebula V3.2 onwards).

X.509 plugins available for command line and for web services authentication.

In continued discussion with Open Grid Forum and those who want to do X.509 Authentication in OpenStack – trying to establish a standard.

X.509 Authentication—how it works

Command line:

- User creates a X.509-based token using "oneuser login" command
- This makes a base64 hash of the user's proxy and certificate chain, combined with a username:expiration date, signed with the user's private key

Web Services:

- Web services daemon contacts OpenNebula XML-RPC core on the users' behalf, using the host certificate to sign the authentication token.
- Use Apache mod_ssl or gLite's GridSite to pass the grid certificate DN (and optionally FQAN) to web services.

Known Current Limitations:

- With Web services, one DN can map to only one user.
- See next talk for X.509 Authorization plans

Security: Incident Response

FermiCloud does not accept that the third party VM signing advocated by HEPiX Virtualisation Working Group is sufficient.

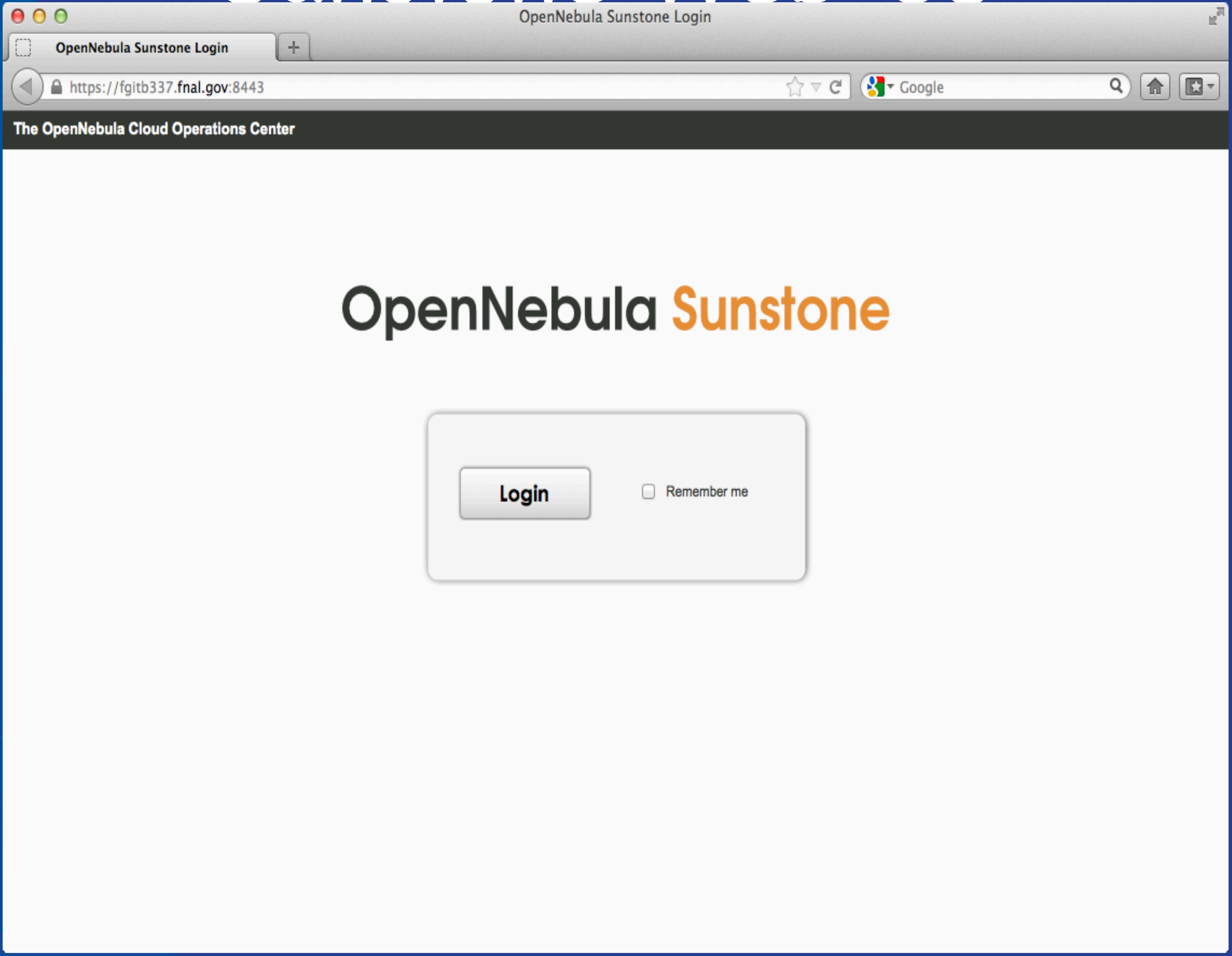
In the event of an incident on FermiCloud, the actions of the following individuals will be examined:

- The individual who transferred the VM image to FermiCloud (yes we have logs).
- The individual who requested that the VM image be launched on FermiCloud (yes we have logs).

If neither individual can provide an acceptable answer, then both individuals may lose FermiCloud access privileges.

Note that virtualization allows us to snapshot memory and process stack of a running VM and capture all evidence.

Sunstone Web UI



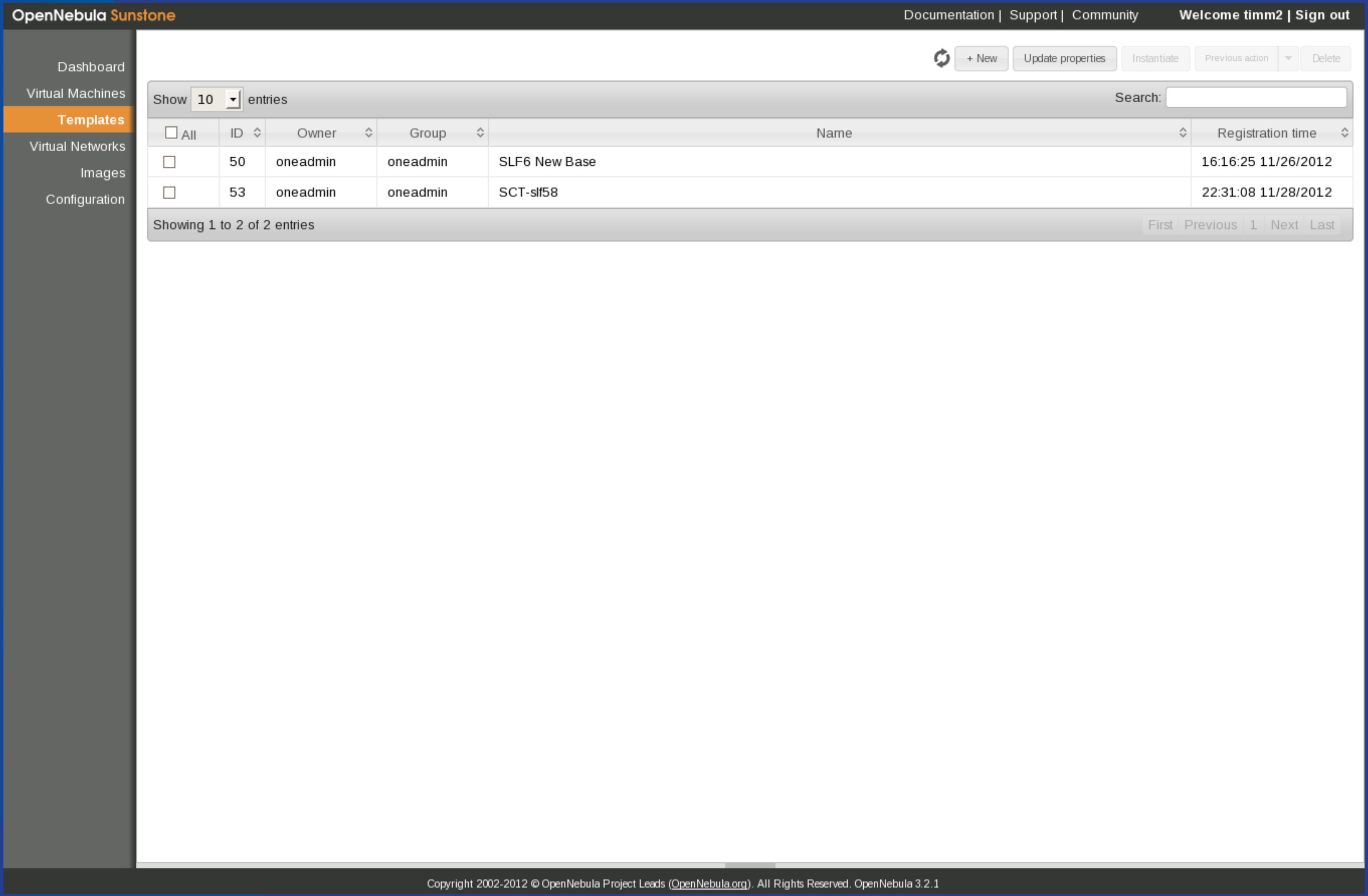
OpenNebula Sunstone Login

https://fgitb337.fnal.gov:8443

The OpenNebula Cloud Operations Center

OpenNebula Sunstone

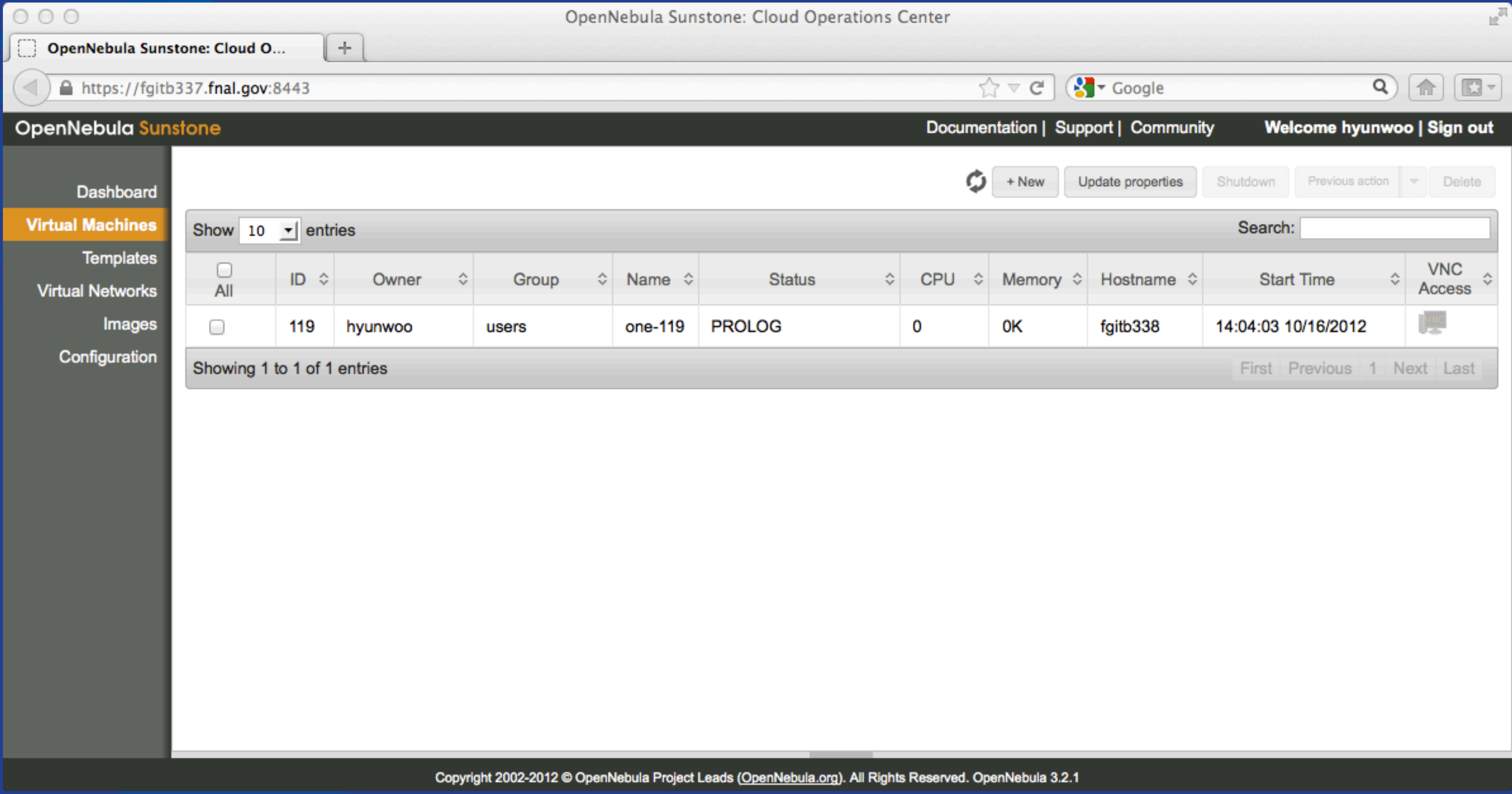
Selecting a template



□ All	ID	Owner	Group	Name	Registration time
□	50	oneadmin	oneadmin	SLF6 New Base	16:16:25 11/26/2012
□	53	oneadmin	oneadmin	SCT-slf58	22:31:08 11/28/2012

Showing 1 to 2 of 2 entries

Launching the Virtual Machine



OpenNebula Sunstone: Cloud Operations Center
OpenNebula Sunstone: Cloud O...
https://fgitb337.fnal.gov:8443
Google
OpenNebula Sunstone
Documentation | Support | Community
Welcome hyunwoo | Sign out
Dashboard
Virtual Machines
Templates
Virtual Networks
Images
Configuration
+ New
Update properties
Shutdown
Previous action
Delete

Show 10 entries

Search:

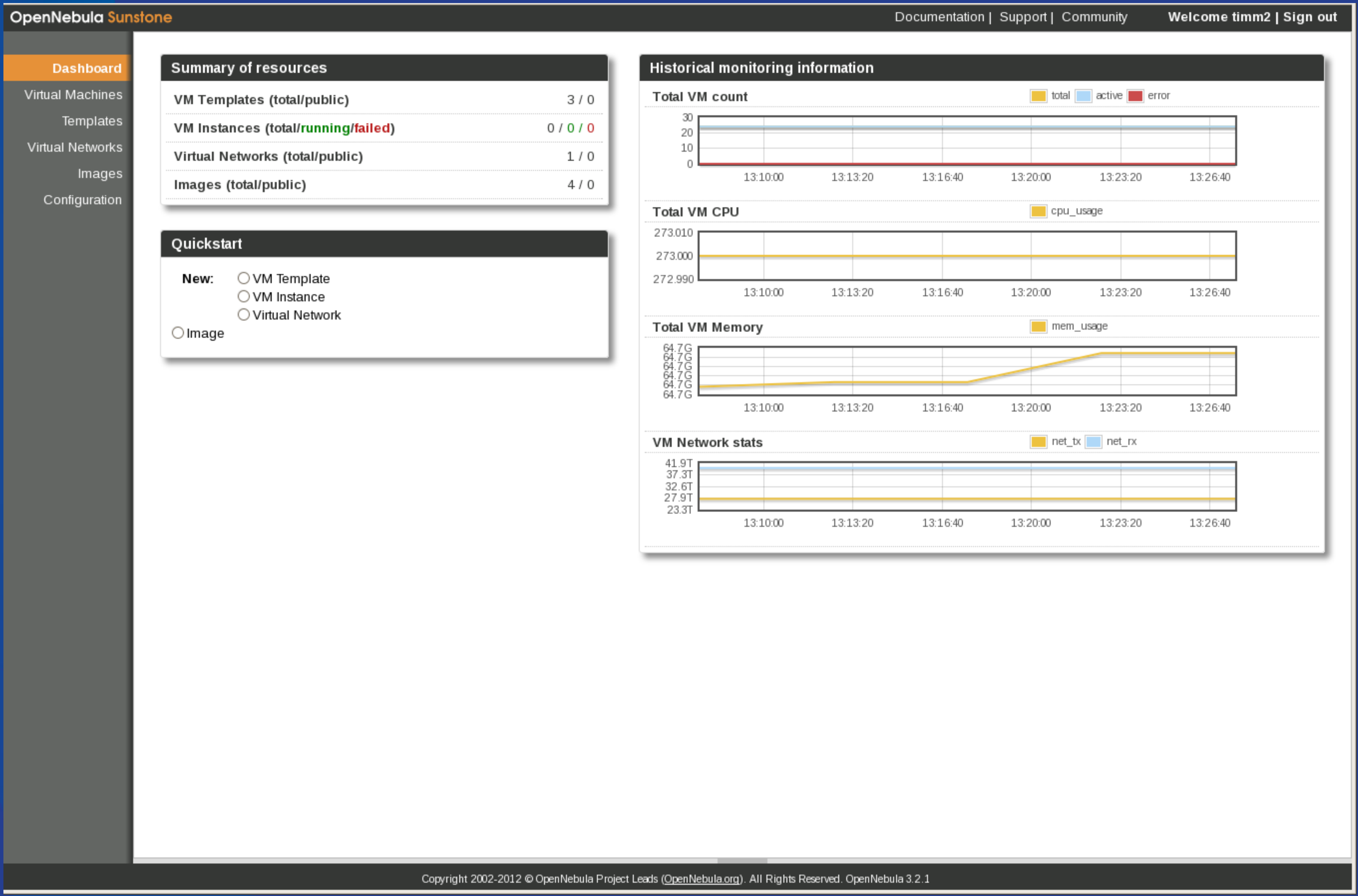
□ All	ID	Owner	Group	Name	Status	CPU	Memory	Hostname	Start Time	VNC Access
□	119	hyunwoo	users	one-119	PROLOG	0	0K	fgitb338	14:04:03 10/16/2012	VNC

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

Copyright 2002-2012 © OpenNebula Project Leads (OpenNebula.org). All Rights Reserved. OpenNebula 3.2.1

Monitoring VM's



Summary of resources

VM Templates (total/public)	3 / 0
VM Instances (total/running/failed)	0 / 0 / 0
Virtual Networks (total/public)	1 / 0
Images (total/public)	4 / 0

Quickstart

New:
○ VM Template
○ VM Instance
○ Virtual Network
○ Image

Historical monitoring information

Copyright 2002-2012 © OpenNebula Project Leads (OpenNebula.org). All Rights Reserved. OpenNebula 3.2.1

Auxiliary Services

- Monitoring
  - Usage monitoring, Nagios, Ganglia, RSV
- Accounting/Billing
  - Gratia
- Installation/Provisioning
  - Cobbler, Puppet
- Web Server
- Secure secrets repositories
- Syslog forwarder
- NIS servers
- Dual MySQL database server (OpenNebula backend)
- LVS frontend.